National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service

Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

## NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

## AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

NL-339 (r. 88/04) c

Canada

National Library of Canada
Canadian Theses Service
Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Service des thèses canadiennes

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-55437-1

Canadä

THE UNIVERSITY OF ALBERTA

# Numerical Techniques for Large Scale Computer Simulation

BY ©

**Hong Jiang**

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

IN

APPLIED MATHEMATICS

DEPARTMENT OF MATHEMATICS
EDMONTON, ALBERTA
FALL 1989

# THE UNIVERSITY OF ALBERTA

## RELEASE FORM

NAME OF AUTHOR: **Hong Jiang**

TITLE OF THESIS:

**Numerical Techniques for Large Scale Computer Simulation**

DEGREE FOR WHICH THESIS WAS PRESENTED: **Doctor of Philosophy**

YEAR THIS DEGREE GRANTED: **1989**

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(Signed) H. Jiang

Permanent Address:
Department of Mathematics
University of Alberta
Edmonton, Alberta
Canada T6G 2G1

Date: May 15, 1989

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled

**Numerical Techniques for Large Scale Computer Simulation**

submitted by **Hong Jiang** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Applied Mathematics**.

Supervisor

H. I. Freedman

W. Allegretto

External Examiner

Date: April 27, 1989

To My Parents and Grandparents

# ABSTRACT

In this thesis, we examine two aspects of numerical simulation of many physical processes, namely, construction of numerical boundary conditions for hyperbolic equations and design of iterative algorithms for solving large linear systems of equations.

In part I, we develop a nonreflecting principle for hyperbolic initial boundary value problems. This principle leads to a pseudodifferential operator which yields a perfectly absorbing boundary condition. Since the perfectly absorbing boundary condition is nonlocal in space and time, it must be approximated by boundary conditions given in terms of differential operators. We show that optimal approximations lead to absorbing boundary conditions in a canonical form which is a product of some lower-order differential operators. Each factor in the canonical form annihilates exactly a wave packet travelling with a certain group velocity.

In part II, we propose a polynomial preconditioning for the conjugate gradient method for solving a system of the form $A^2u = b$, and analyze the convergence behavior of this preconditioning algorithm. We show analytically that the new algorithm yields a convergence rate comparable to a polynomial preconditioning algorithm applied to the better conditioned system $Au = b$. Finding an optimal polynomial for the new algorithm is tantamount to solving a constrained minimax problem. We develop a method which efficiently solves the constrained minimax problem.

# ACKNOWLEDGEMENTS

With my deepest appreciation, I wish to thank my supervisor Dr. Yau Shu Wong for his invaluable guidance and friendly support.

I also wish to thank Drs. Walter Allegretto, Herbert I. Freedman, S. Sivaloganathan and Robert J. Tait for their advice and interest in my research.

A special thanks goes to N. Sivakumar and Dr. S. Sivaloganathan for their careful proofreading of this thesis.

I thank the Department of Mathematics for providing congenial research environment. I am grateful to Dr. Yau Shu Wong, the Killam Scholarship Committee and the University of Alberta for their financial support.

# TABLE OF CONTENTS

CHAPTER PAGE

1. INTRODUCTION 1

1.1 Artificial Boundary Conditions 2

1.2 Preconditioned Conjugate Gradient Method 4

1.3 Outline of the Thesis 6

2. INITIAL BOUNDARY VALUE PROBLEMS 10

2.1 Well-posedness of Initial Boundary Value Problems 10

2.2 Some Assumptions 15

2.3 Absorption Properties of Boundary Operators 19

3. WAVE PROPAGATION IN INITIAL BOUNDARY VALUE PROBLEMS 30

3.1 Group Velocity 30

3.2 Leftgoing and Rightgoing Solutions 33

3.3 Strongly Well-posedness 36

3.4 Reflection Coefficients 41

4. ABSORBING BOUNDARY CONDITIONS 43

4.1 Perfectly Absorbing Boundary Conditions 43

4.2 Rational Approximations 49

4.3 Characterization of the Best Absorbing Operators 54

4.4 Selection of Parameters 60

5. SOME SECOND-ORDER HYPERBOLIC EQUATIONS 63

5.1 The Wave Equation and the Klein-Gordon Equation 63

5.2 Transonic Small Disturbance Equation 66

5.3 Numerical Computations 72

6. ABSORBING BOUNDARY CONDITIONS FOR FINITE DIFFERENCE SCHEMES 85
6.1 General Consideration 87
6.2 Absorbing Boundary Conditions 90
6.3 Examples 94

7. POLYNOMIAL PRECONDITIONING FOR CONJUGATE GRADIENT METHODS 110
7.1 Preconditioned Conjugate Gradient Algorithm 110
7.2 Polynomial Preconditionings 116

8. A POLYNOMIAL PRECONDITIONING METHOD APPLIED TO THE SYSTEM $A^2u = b$ 122
8.1 The Approximation Polynomial Preconditioning 122
8.2 The Minimax Polynomial 124
8.3 The Least Squares Polynomial 133

9. COMPUTATION OF THE MINIMAX POLYNOMIAL 137
9.1 Introduction 137
9.2 Characterization of $R_{m+2}$ 139
9.3 An Algorithm for Computation of $R_{m+2}$ 145

10. NUMERICAL EXPERIMENTS 152
10.1 A Model Problem 153
10.2 The Biharmonic Equation 155

11. CONCLUSION AND DISCUSSION 160

REFERENCES 165

# LIST OF FIGURES

| FIGURE | | PAGE |
|---|---|---|
| 1.0.1 | The computational area | 2 |
| 3.1.1 | Group Velocity | 33 |
| 3.3.1 | Weak ill-posedness | 38 |
| 5.3.1 | The enrgy of solution | 79 |
| 5.3.2 | The enrgy of solutions | 80 |
| 5.3.3 | The enrgy of solutions | 80 |
| 5.3.4 | The enrgy of solutions | 81 |
| 5.3.5 | Large computational domain | 81 |
| 5.3.6 | The enrgy of reflected waves: Upstream | 83 |
| 5.3.7 | The enrgy of reflected waves: Downstream | 83 |
| 5.3.8 | The enrgy of reflected waves: Sidewall( $y = 1$ ) | 84 |
| 6.1.1 | Dispersion relation of LF, $\lambda < 1$ | 89 |
| 6.1.2 | Parasitic waves | 90 |
| 6.3.1 | Solutions of LF | 98 |
| 6.3.2 | Solutions of LF | 98 |
| 6.3.3 | Dispersion relation | 99 |
| 6.3.4 | Total reflection | 101 |
| 6.3.5 | Reflections of first-order boundary conditions | 102 |
| 6.3.6 | Reflections of second-order boundary conditions | 104 |
| 6.3.7 | Reflections of third-order boundary conditions | 105 |
| 6.3.8 | Total absorption | 106 |
| 6.3.9 | Effects of generalized eigenvalues | 108 |
| 9.2.1 | The function $\alpha = \alpha_1(\beta)$ | 144 |
| 9.2.2 | The function $\alpha = \alpha_2(\beta)$ | 145 |

9.3.1 The residual polynomial $R_{m+2}$ 149

9.3.2 The minimax preconditioning polynomial $q_m$ 150

10.2.1 CPU time 157

10.2.2 CPU time 159

# LIST OF TABLES

TABLE PAGE

10.1.1 Number of iterations and CPU time 154

10.1.2 Comparision of numbers of iterations 154

10.2.1 Number of iterations for $n = 9801$ 156

10.2.2 Number of iterations 158

## LIST OF DEFINITIONS

| DEFINITION | PAGE |
|---|---|
| Def.2.1.3 | 13 |
| Def.2.1.4 | 14 |
| Def.2.2.4 | 17 |
| Def.2.3.1 | 20 |
| Def.2.3.4 | 22 |
| Def.3.1.1 | 31 |
| Def.3.2.2 | 36 |
| Def.3.3.2 | 40 |
| Def.3.3.3 | 40 |
| Def.3.4.1 | 42 |
| Def.4.2.1 | 51 |
| Def.4.2.3 | 52 |
| Def.4.2.4 | 53 |
| Def.7.1.6 | 115 |
| Def.8.2.1 | 125 |

# LIST OF THEOREMS

| THEOREM | PAGE |
|---|---|
| Thm.2.1.5 | 14 |
| Thm.2.3.2 | 20 |
| Thm.2.3.3 | 21 |
| Thm.2.3.5 | 22 |
| Thm.2.3.6 | 24 |
| Thm.3.2.1 | 35 |
| Thm.4.1.1 | 44 |
| Thm.4.1.2 | 47 |
| Thm.4.2.5 | 53 |
| Thm.4.3.1 | 54 |
| Thm.4.3.2 | 56 |
| Thm.4.3.3 | 56 |
| Thm.4.3.4 | 58 |
| Thm.4.3.5 | 59 |
| Thm.6.2.1 | 92 |
| Thm.7.1.1 | 113 |
| Thm.7.1.5 | 114 |
| Thm.7.2.1 | 117 |
| Thm.7.2.2 | 118 |
| Thm.7.2.3 | 119 |
| Thm.7.2.4 | 120 |
| Thm.7.2.5 | 121 |
| Thm.8.2.7 | 130 |
| Thm.8.2.9 | 132 |

| | |
|---|---|
| Thm.8.3.1 | 134 |
| Thm.9.2.1 | 140 |
| Thm.9.2.2 | 140 |
| Thm.9.2.6 | 144 |
| Thm.9.3.2 | 148 |

# INDEX OF NOTATIONS

The page number denotes the page where the definition is given.

| NOTATION | PAGE |
|---|---|
| $\mathcal{A}_0$ | 17 |
| $\bar{\mathcal{A}}_0$ | 17 |
| $\mathcal{A}_{\tau_0}$ | 17 |
| $\mathring{C}_x$ | 34 |
| $E(\tau,\eta)$ | 12 |
| $E^+(\tau,\eta)$ | 13 |
| $\mathcal{L}^\infty_\epsilon$ | 53 |
| $\mathcal{L}^1$ | 53 |
| $\mathcal{L}^2$ | 53 |
| $N^+_B(\tau,\eta)$ | 14 |
| $N^-_B(\tau,\eta)$ | 23 |
| $\mathcal{O}$ | 23 |
| $\mathring{Q}(\tau,\xi,\eta)$ | 44 |
| $Q(\tau,\xi,\eta)$ | 46 |
| $Q^-(\tau,\eta)$ | 47 |
| $R_{m+2}$ | 125 |
| $\tilde{R}_{m+2}$ | 133 |
| $T_m(x)$ | 115 |
| $\mathcal{V}$ | 31 |
| $\Gamma$ | 31 |
| $\Gamma_0$ | 52 |
| $\Gamma_\epsilon$ | 52 |

| | |
|---|---|
| $\Delta_B^+(\tau,\eta)$ | 14 |
| $\theta_m$ | 125 |
| $\tilde{\theta}_m$ | 133 |
| $\nu$ | 115 |
| $\nu_m$ | 120 |
| $\tilde{\nu}_m$ | 121 |
| $\nu_{APP(m)}$ | 126 |
| $\tilde{\nu}_{APP(m)}$ | 134 |
| $\nu_{TPP(m)}$ | 130 |
| $\Pi_m$ | 52 |
| $\Pi^1$ | 112 |
| $\Pi^{1,0}$ | 125 |

# CHAPTER 1

# INTRODUCTION

Numerical simulations of many physical processes often require calculation of solutions to partial differential equations on some infinite domain. The important areas of application include meteorology [Roac72], seismology [KATW74] and transonic aerodynamics [BaGo77]. The numerical solutions of such problems involve two essential approximations: first, the unbounded physical domain must be approximated by a finite computational domain; and second, the continuum solution must be approximated by some grid solution on a discrete set, and in general, the grid solution is obtained by solving systems of linear equations. Thus a successful numerical simulation process depends on the development of accurate computational boundary conditions and efficient algorithms for solving large systems of equations. These are the themes of this thesis.

The work in this thesis was motivated by problems of numerical simulation of unsteady transonic flows, which are important in many applications such as flutter analysis, helicopter blades and manoeuvering flight. The prototype of such problems is the flow over an oscillating airfoil as depicted in Fig.1.0.1. On the computational walls, boundary conditions are needed so that the artificial wall effect created by the computational walls is minimized.

After correct boundary conditions are found, the governing equation of the flow field can be discretized. The solution of the nonlinear equation is obtained, after some linearization process, via solving a sequence of linear systems of algebraic equations. Typically, for a three dimensional problem, the matrix of the system is large and sparse. Its solution requires an efficient iterative algorithm.

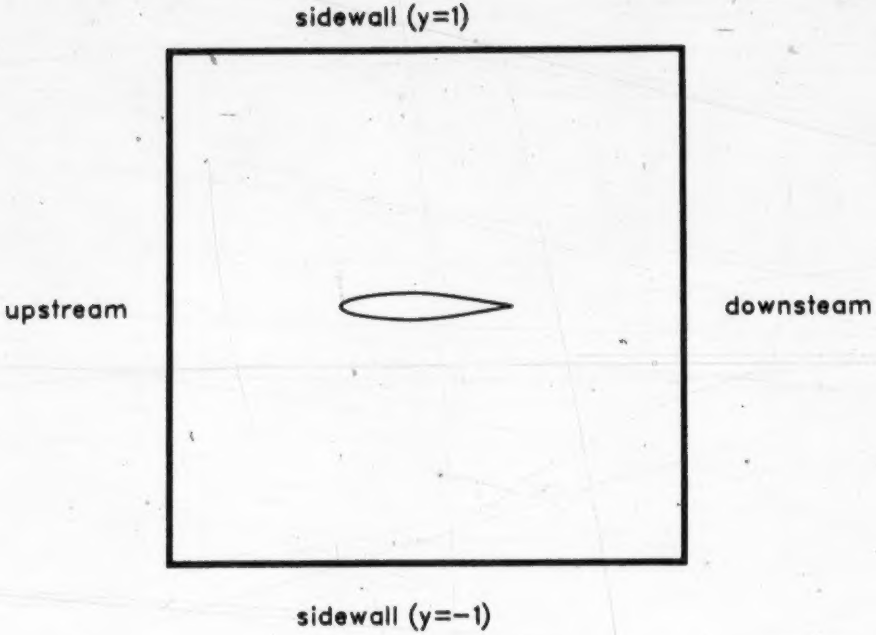

**Figure 1.0.1** The computational area

In this thesis, we attempt to address these two issues, namely, the construction of artificial boundary conditions that minimizes the unphysical effect of the artificial boundaries, and the design of an efficient iterative algorithm for solutions of large sparse matrices. Although the example of transonic flows motivated our work, the scope of this thesis is not limited to the calculation of transonic flows. The techniques discussed in this thesis can well be applied to large scale numerical simulations for many problems in science and engineering.

## 1.1 Artificial Boundary Conditions

For steady state problems, coordinate transform techniques have proved to be an efficient way of handling computational boundaries [GrOr77]. The success of the coordinate transform lies in the fact that the solutions of steady state problems behave in a simple way at infinity; either they vanish rapidly or approach a constant value at infinity. However, the solutions of genuinely unsteady state problems often possess strong oscillatory transient behaviors at infinity, and such solutions are not amenable to the coordinate transform technique. Under such conditions an alternative procedure for treating the computational boundaries is essential.

When artificial boundaries are introduced to limit the area of computation, some boundary conditions are needed at these artificial boundaries in order to guarantee a unique and well-posed problem. In general, the presence of an artificial boundary may result in a solution which is different from the free space solution. This is because the hitting of the boundary by a wave-like disturbance from the interior causes reflection to occur, generating waves coming into the domain. This reflection is totally unphysical, and reduces the accuracy of the solution. It is therefore desirable to develop some boundary conditions that allow waves to propagate through the artificial boundary as if there were no boundary present. Boundary conditions that possess such a property is referred to as *absorbing boundary conditions*.

Absorbing boundary conditions have been of substantial interest in recent computational literature. In [Smit74] and [Lind75] some special methods were used to derive absorbing boundary conditions for the wave equation. Engquist and Majda [EnMa77] [EnMa79] proposed some parabolic wave equations (or one-way wave equations) as absorbing boundary conditions. (See also [TrHa86], [BEHJ88a] and [BEHJ88b]). Bayliss and Turkel [BaTu80] used the asymptotic expansion of solutions in the far field to obtain similar boundary conditions. Gustafsson and Kreiss [GuKr79] considered a hyperbolic system of equations for two functions of $x, y$ and $t$ in a waveguide. Gustafsson [Gust88] obtained inhomogeneous boundary conditions. Hedstrom [Heds79], Thompson [Thom87] and Hagstrom and Hariharan [HaHa88] considered some nonlinear hyperbolic systems. Wilson [Wils82], Wagatha [Waga85] and Higdon [Higd86a] [Higd87] constructed absorbing boundary conditions for finite difference approximations to the wave equation.

In this thesis, we consider a physical process which is governed by a linear hyperbolic equation with constant coefficients in an infinite domain, say, $x \in \mathbb{R}$ . This is equivalent to a purely initial value problem. We look for a boundary condition at $x = 0$ such that the solution of the purely initial value problem can be well approximated by that of the initial boundary value problem in the domain $x \geq 0$ . We will apply the theory of wave propagation to develop a nonreflecting principle, from which perfectly absorbing boundary conditions can be derived. These perfectly absorbing boundary conditions are given by some pseudodifferential operators which are nonlocal in both space and time; they are not numerically useful in practice. The use of some rational approximations leads to local, highly absorbing boundary conditions. We will also show that the optimal boundary conditions must be in a canonical form.

## 1.2 Preconditioned Conjugate Gradient Method

Once the mathematical equations and boundary conditions are formulated, the final stage in a numerical simulation process often involves solving a large sparse linear system $Au = b$ . Obtaining the solution for a large system of linear equations is frequently the most time consuming part in the entire simulation procedure. When $A$ is symmetric and positive definite, the conjugate gradient (CG) method of Hestenes and Stiefel [HeSt52] is often used. Unfortunately, the CG method may converge slowly if $A$ is ill-conditioned. In such cases, a preconditioned conjugate gradient (PCG) method [CGO76] is needed. Let $M = QQ^T$ , where $Q$ is a nonsingular matrix, be a preconditioner. Then the PCG method is in fact the CG method applied to the system $Q^{-1}AQ^{-T}Q^{T}u = Q^{-1}b$ . This system has an effective condition number equal to that of the matrix $M^{-1}A$ . Thus by constricting the eigenvalues of $M^{-1}A$ , a good preconditioner $M$ speeds up convergence. A popular preconditioning is based on an $LU$ incomplete factorization procedure [MevV77] [Gust78]. In this procedure, backward- and forward-solving of triangular matrices are required for the calculation of $M^{-1}$ , and the process is inherently scalar, not suitable to vector or parallel computers. Dubois *et al.* [DGR79] proposed a preconditioning where the incomplete factorization is replaced by an incomplete inverse. The incomplete inverse is in fact a polynomial in $A$ , i.e., $M^{-1} = p(A)$ for some polynomial $p$ .

Polynomial preconditioning is attractive for the following two reasons. First, it is quite simple to adopt. Only matrix by vector multiplications are involved in calculations and no extra storage is needed for $M$ (or $M^{-1}$ ). The second advantage of polynomial preconditioning is that it can be efficiently implemented on vector or parallel computers. In fact, Chan *et al.* [CKT88] observed that on some parallel computers, polynomial preconditioning performs better than most other preconditioning methods.

In this thesis, we will investigate the polynomial preconditioning technique applied to large linear systems of equations. By applying an approximate polynomial preconditioning technique, we will show that the solution of this system can be obtained much more efficiently than when using the traditional polynomial method. The basic idea of an approximate polynomial preconditioning is that the incomplete inverse is obtained from a polynomial based on another coefficient matrix that consists of far fewer non-zero entries than the original matrix operator. By contrast, the traditional polynomial preconditioner is formed from a polynomial using the original coefficient matrix. We will also illustrate how this new preconditioning technique can be used in many problems of physical interest.

## 1.3 Outline of the Thesis

This thesis is divided into two parts; each part is devoted to one of the themes discussed above. Chapters two through six are concerned with the development of absorbing boundary conditions, and chapters seven through ten deal with the conjugate gradient method with polynomial preconditioning.

In chapter two, we consider general hyperbolic initial boundary value problems. This chapter serves as a theoretical framework for the development of subsequent chapters. In the discussion, the artificial boundary is located at $x = 0$ , and the solution is sought in the half space $x \geq 0$ . We first review the theory of well-posed hyperbolic initial boundary value problems in sections one and two. Then in section three, we prove a theorem which provides a criterion for a boundary condition to be absorbing.

Chapter three is a survey of the theory of wave propagation in hyperbolic initial boundary value problems. We also introduce a speed of translation of amplitude, which was first used by Trefethen [Tref82] in finite difference schemes.

Chapter four includes the main contributions of part I of this thesis. In this chapter, we develop a nonreflecting principle and prove a characterization theorem which states that the optimal local absorbing boundary conditions are always representable in a certain canonical form.

In chapter five, we apply the theory of chapter four to some equations of physical interest, and show that our theory results in simplications and generalizations of boundary conditions derived earlier for these examples by Engquist and Majda [EnMa80].

In chapter six, we derive absorbing boundary conditions for finite difference equations resulting from the discretization of hyperbolic differential equations. Some numerical results are also presented.

In chapter seven, we review the preconditioned conjugate gradient algorithm and its convergence behavior, in particular, with polynomial preconditionings.

In chapter eight, we propose a preconditioning method for the solution of linear systems of equations, and show that this new preconditioning is superior to the traditional ones.

In chapter nine, the minimax polynomial resulting from the preconditioning of chapter eight is studied. We provide a new algorithm to compute the coefficients of this polynomial.

In chapter ten, some numerical results on the new preconditioning method are reported.

In chapter eleven, some final discussions are given.

Each chapter is divided into sections. When a reference is made to a section, the section number is prefixed by §, e.g., § 1.3 refers to the third section in chapter 1 Lemmas, theorems, corollaries and definitions are listed as items in each section. Their reference will be abbreviated by Lem., Thm., Cor. and Def. respectively. Thus Thm.2.3.6 means the theorem that appears in section 3 of chapter 2 as item 6 (not necessarily the 6th theorem). Equations, figures and tables are numbered similarly.

The first two sections of chapter two, the entire chapter three, the first section of chapter six and chapter seven contain mostly survey results. § 5.3, § 6.3 and chapter ten are concerned with computations and numerical results. The rest deals with theoretical analysis.

Some results of this thesis have appeared in [JiWo89] and [WoJi89].

# PART I ABSORBING BOUNDARY CONDITIONS

# CHAPTER 2

# INITIAL BOUNDARY VALUE PROBLEMS

This chapter deals with some basic theory of initial boundary value problems. The technique used is the Fourier-Laplace transform (see for example [Pete83] and [Saka82]). This chapter serves as a theoretical foundation for subsequent developments in the later chapters. In section one, we review the theory of well-posedness of the initial boundary value problem developed by Hersh [Hers64]. Some special cases of initial boundary value problems are considered in section two. In section three, relations between purely initial value problems and initial boundary value problems are investigated. The concept of absorbing boundary conditions is introduced. In particular, we give a necessary and sufficient condition for the solution of the two problems to be identical.

## 2.1 Well-posedness of Initial Boundary Value Problems

We consider the initial boundary value problem (IBVP):

$$P(D_t, D_x, D_y)u = 0\,, \qquad 0 < t, x < +\infty, \quad y \in \mathbb{R}^{n-1} \tag{2.1.1a}$$

$$B^k(D_t, D_x, D_y)u = g^k\,, \qquad x = 0, \quad k = 1, 2, \ldots, r \tag{2.1.1b}$$

$$\frac{\partial^k}{\partial t^k}u = f^k\,, \qquad t = 0, \quad k = 0, 1, \ldots, p-1\,. \tag{2.1.1c}$$

In the above equations [1] , $P(\tau, \xi, \eta)$ is a real polynomial of degree $p$ in the variable $(\tau, \xi, \eta) \in \mathbb{C} \times \mathbb{C} \times \mathbb{C}^{n-1}$ . And $D_t = \frac{1}{i}\frac{\partial}{\partial t}, D_x = \frac{1}{i}\frac{\partial}{\partial x}$ and $D_y = (\frac{1}{i}\frac{\partial}{\partial y_1}, \ldots, \frac{1}{i}\frac{\partial}{\partial y_{n-1}})$ . We will denote the data $\{f^k\}, \{g^k\}$ in vector form $f$, $g$ , respectively.

---

[1] We will denote the dual of $(t, x, y)$ by $(\tau, \xi, \eta)$ .

We assume that the equation Eq.(2.1.1a) is hyperbolic in the direction of $t$, that is, it satisfies the following Gårding condition.

2.1.1 Gårding CONDITIONS [Gård50].

*(1) there exists a* $\tau_0 > 0$ *such that*

$$P(\tau, \xi, \eta) \neq 0, \qquad \text{for } Im\,\tau < -\tau_0 \text{ and } (\xi, \eta) \in \mathbb{R}^n,$$

*(2)* $P_0(1,0,0) \neq 0$, *where* $P_0$ *is the principal part of* $P$.

Gårding conditions imply that the purely initial value problem is well-posed [Shil68].

The existence and uniqueness of the solution of the IBVP Eq.(2.1.1) have been established under very general conditions. In particular, Hersh [Hers64] considered square integrable solutions of Eq.(2.1.1). Sakamoto [Saka82] studied infinitely differentiable solutions. Kreiss [Krei70] studied first-order hyperbolic systems. We will briefly review the results of Hersh in this section. The results of Kreiss will be discussed in the next chapter.

For a solution of Eq.(2.1.1) to exist, it is necessary that $f, g$ satisfy some compatibility conditions. Detailed discussions of these conditions can be found in [Saka82]. When such conditions are satisfied, we will say that $f$, $g$ are admissible data.

Let $H$ be a solution space, which may be some *Sobolev space* (for integrable solutions) or a *Fréchet space* (for infinitely differentiable solutions). The IBVP Eq.(2.1.1) is well-posed if it admits a unique solution for arbitrary admissible data $f, g$. Furthermore, the solution is continuously dependent on the prescribed data

$f, g$ . Typically, the solution must satisfy an estimate

$$\|u(t)\|_{\Omega} \leq C\Big\{\|f\|_{\Omega} + \|g\|_{\partial\Omega\times[0,T]}\Big\}, \tag{2.1.2}$$

where $\Omega = [0, +\infty)$ , and $C = C(t)$ is independent of $f, g$ .

To study the solutions of Eq.(2.1.1), we first consider its special solutions of exponential form

$$u = e^{i(\tau t + \eta y)}\phi(x), \tag{2.1.3}$$

where $(\tau, \eta) \in \mathbb{C} \times \mathbb{R}^{n-1}$ , with $\operatorname{Im}\tau < -\tau_0$ . When the solution Eq.(2.1.3) is substituted into Eq.(2.1.1a), the result is

$$P(\tau, D_x, \eta)\phi = 0, \qquad x > 0. \tag{2.1.4}$$

We are interested in the solutions of Eq.(2.1.4) which are bounded in the domain $[0, +\infty)$ . This problem is said to be the *eigenvalue problem* associated with the IBVP Eq.(2.1.1).

For a given $(\tau, \eta) \in \mathbb{C} \times \mathbb{R}^{n-1}$ , with $\operatorname{Im}\tau < -\tau_0$ , Eq.(2.1.4) is a linear ordinary differential equation. Its solution space will be denoted by $E(\tau, \eta)$ . Then for each $\phi \in E(\tau, \eta)$ , $\phi$ can be expressed as

$$\phi = \sum_{j,k} c^{+}_{j,k} x^{k} e^{i\xi^{+}_{j} x} + \sum_{j,k} c^{-}_{j,k} x^{k} e^{i\xi^{-}_{j} x}.$$

In the above expression, $\xi^{+}_{j}$, $\xi^{-}_{j}$ are the roots of $P(\tau, \xi, \eta) = 0$ for fixed $(\tau, \eta) \in \mathbb{C} \times \mathbb{R}^{n-1}$ . They are classified into two groups

$$\begin{aligned} \xi^{+}_{j} &= \xi^{+}_{j}(\tau, \eta), \qquad \operatorname{Im}\xi^{+}_{j} > 0, \quad j = 1, \dots, m^{+} \\ \xi^{-}_{j} &= \xi^{-}_{j}(\tau, \eta), \qquad \operatorname{Im}\xi^{-}_{j} < 0, \quad j = 1, \dots, m^{-}. \end{aligned} \tag{2.1.5}$$

where $m^{\pm} = m^{\pm}(\tau, \eta)$ is the number of the roots, counted according to their multiplicities, with the positive and the negative imaginary parts, respectively.

2.1.2 LEMMA [Hers64]. *For* $Im\,\tau < -\tau_0$ *and* $\eta$ *real,* $m^{\pm}$ *are constants.*

PROOF: $m^{\pm}$ are continuous functions of $\tau, \eta$ ; and the values of $m^{\pm}$ will only change if there is a root $\xi$ with zero imaginary part. However, according to (1) of the Gårding condition, there exists no real root $\xi$ if $\mathrm{Im}\,\tau < -\tau_0$ , and $\eta \in \mathbb{R}^{n-1}$ . This shows that $m^{\pm}$ must remain constant. □

With the classification in Eq.(2.1.5), the solution space $E(\tau,\eta)$ can be decomposed as

$$E(\tau,\eta) = E^{+}(\tau,\eta) \oplus E^{-}(\tau,\eta),$$

where $E^{\pm}(\tau,\eta)$ of dimension $m^{\pm}$ is the space of the solutions $\phi$ corresponding to $\xi^{\pm}$ . It is clear that a solution $\phi \in E^{+}(\tau,\eta)$ decays exponentially for $x \in [0,+\infty)$ . If such a solution also satisfies the homogeneous boundary condition

$$B^{k}(\tau, D_x, \eta)\phi = 0, \quad x = 0,\ k = 1, ldots, r, \tag{2.1.6}$$

then the IBVP Eq.(2.1.1) cannot be well-posed because a family of solutions $u_\alpha$ of the form Eq.(2.1.3) can be constructed so that $\|u_\alpha(t)\|_\Omega$ grows exponentially in $\alpha$ when $t > 0$ , while the initial data remain bounded at $t = 0$ . That is, the solutions do not depend continuously on the initial data.

2.1.3 DEFINITION. *Given* $(\tau,\eta) \in \mathbb{C} \times \mathbb{R}^{n-1}$ , *with* $Im\,\tau < -\tau_0$ , $\phi$ *is said to be an eigenfunction of Eq.(2.1.1) if* $\phi \neq 0$ , *and*

*(1)* $\phi \in E^{+}(\tau,\eta)$ ,

*(2)* $B^{k}(\tau, D_x, \eta)\phi = 0, \qquad k = 1, \ldots, r$ .

Therefore, the well-posedness of Eq.(2.1.1) implies nonexistence of eigenfunctions. If $\{\phi_1, \ldots, \phi_{m^+}\}$ is a basis for $E^{+}(\tau,\eta)$ , then the eigenfunctions are linear combinations of these $\phi_j$ 's:

$$\phi = \sum_{j=1}^{m^+} c_j \phi_j \, .$$

If solutions of the above form are inserted into the homogeneous boundary condition Eq.(2.1.6), the result is

$$B[\phi_1(0), \dots, \phi_{m^+}(0)]c = 0 \, , \tag{2.1.7}$$

where $c = [c_1, \dots, c_{m^+}]^T$ , and $B[\phi_1(0), \dots, \phi_{m^+}(0)]$ is an $m^+ \times m^+$ matrix. A nonzero solution $c$ of Eq.(2.1.7) corresponds to an eigenfunction; therefore when the $m^+ \times m^+$ matrix

$$N_B^+(\tau, \eta) = B[\phi_1(0), \dots, \phi_{m^+}(0)] \tag{2.1.8}$$

becomes singular, there is an eigenfunction of Eq.(2.1.1).

2.1.4 DEFINITION. *Let* $(\tau, \eta) \in \mathbb{C} \times \mathbb{R}^{n-1}$ , *with* $\mathrm{Im}\, \tau < -\tau_0$ . $(\tau, \eta)$ *is said to be an eigenvalue of Eq.(2.1.1) if*

$$\Delta_B^+(\tau, \eta) = \det N_B^+(\tau, \eta) = 0 \, .$$

Thus $\Delta_B^+ \neq 0$ for all $(\tau, \eta) \in \mathbb{C} \times \mathbb{R}^{n-1}$ , with $\mathrm{Im}\, \tau < -\tau_0$ is a necessary condition for Eq.(2.1.1) to be well-posed. On the other hand, if $\Delta_B^+ \neq 0$ , Hersh [Hers64] was able to construct explicitly a solution of Eq.(2.1.1) by the Fourier-Laplace transform. Therefore, we have

2.1.5 THEOREM [Hers64]. *The IBVP Eq.(2.1.1) is well-posed if and only if there exists a* $\tau_1 > \tau_0$ *such that* $\Delta_B^+(\tau, \eta) \neq 0$ *for all* $\mathrm{Im}\, \tau < -\tau_1$ *and* $\eta \in \mathbb{R}^{n-1}$ .

PROOF: See [Hers64]. □

In Thm.2.1.5, the solution space $H$ in question is the space of tempered distributions. If a tempered distribution $u$ is regular, it is a function of slower growth, i.e., there are some $k$ and $M$ such that $|u(x)| \leq M|x|^k$ as $x \to +\infty$ .

2.1.6 COROLLARY. *A necessary condition for the well-posedness of Eq.(2.1.1) is* $r = m^+$ .

PROOF: If $r < m^+$ , Eq.(2.1.7) admits a nontrival solution $c$ . This leads to an eigenfunction, implying Eq.(2.1.1) is not well-posed. On the other hand, if $r > m^+$ , Eq.(2.1.1) may not have any solution for some nonzero data $g$ . ☐

Sakamoto [Saka82] studied solutions of Eq.(2.1.1) in a Fréchet space, the space of infinitely differentiable functions. She obtained conditions for well-posed IBVP Eq.(2.1.1) similar to that of Thm.2.1.5. In addition Sakamoto also studied equations of variable coefficients.

Kreiss [Krei70] considered first-order hyperbolic systems with variable coefficients. He obtained a necessary and sufficient condition for well-posed initial boundary value problems known as the ***normal mode analysis***. Kreiss' condition, which will be discussed later, is stronger than that of Thm.2.1.5. In obtaining an estimate of the form Eq.(2.1.2), Kreiss used the $\mathcal{L}^2$ norm.

## 2.2 Some Assumptions

In this section, we derive an explicit formula for $\Delta_B^+$ in some special cases.

2.2.1 ASSUMPTION. *For any* $(\tau, \eta) \in \mathbb{C} \times \mathbb{R}^{n-1}$ , *with* $Im\,\tau < -\tau_0$ , *we assume that* $P(\tau, \xi, \eta) = 0$ *has no multiple roots* $\xi$ .

Under this assumption, each root in Eq.(2.1.5) is distinct. Then a basis of

$E^+(\tau,\eta)$ can be chosen as $\{\phi_1,\dots,\phi_{m^+}\}$ where

$$\phi_j = e^{i\xi_j^+ x}, \qquad j = 1,\dots,m^+ .$$

The matrix $N_B^+(\tau,\eta)$ of Eq.(2.1.8) can be found to be

$$[N_B^+(\tau,\eta)]_{k,j} = B^k(\tau,\xi_j^+(\tau,\eta),\eta), \qquad \begin{cases} k = 1,\dots,m^+ \\ j = 1,\dots,m^+ , \end{cases}$$

and

$$\Delta_B^+(\tau,\eta) = \det[B^k(\tau,\xi_j^+(\tau,\eta),\eta)], \tag{2.2.1}$$

where $\{B^k\}$ is the boundary operator of Eq.(2.1.1b).

In particular, if the boundary operator of Eq.(2.1.1b) is chosen as

$$B^k(D_t,D_x,D_y)u = \frac{\partial^{k-1}u}{\partial x^{k-1}},$$

i.e., $B^k(\tau,\xi,\eta) = \xi^{k-1}$, $\qquad k = 1,\dots,m^+$ , then

$$\Delta_B^+ = \prod_{j\neq k}(\xi_j^+ - \xi_k^+).$$

By Assumption 2.2.1, $\xi_j \neq \xi_k$ if $j \neq k$ . Therefore the above expression does not vanish. From this and Thm.2.1.5, one has the following corollary.

2.2.2 COROLLARY. *If $P$ is hyperbolic in the direction of $t$ , then the IBVP*

$$P(D_t,D_x,D_y)u = 0, \qquad 0 < t, x < +\infty, \quad y \in \mathbb{R}^{n-1} \tag{2.2.2a}$$

$$\frac{\partial^k}{\partial x^k}u = g^k, \qquad x = 0, \quad k = 0,1,\dots,m^+ - 1 \tag{2.2.2b}$$

$$\frac{\partial^k}{\partial t^k}u = f^k, \qquad t = 0, \quad k = 0,1,\dots,p-1 \tag{2.2.2c}$$

*is well-posed.*

As another special case, we consider homogeneous equations. We assume that both $P(\tau,\xi,\eta)$ and $B^k(\tau,\xi,\eta)$ are homogeneous polynomials of degree $p$ and $d_k$ respectively. Then the Gårding condition (1) reads

$$(1') \quad P(\tau,\xi,\eta) \neq 0 \text{ if } \operatorname{Im}\tau \neq 0 \text{ and } (\xi,\eta) \in \mathbb{R}^n .$$

In this case, for each $(\tau,\eta) \in \mathbb{C} \times \mathbb{R}^{n-1}$, with $\operatorname{Im}\tau \neq 0$, the classification of the roots of $P(\tau,\xi,\eta) = 0$ in terms of Eq.(2.1.5) is possible. Furthermore, $\xi_j^\pm = \xi_j^\pm(\tau,\eta)$ is a homogeneous function of degree $1$. Consequently, $\Delta_B^+(\tau,\eta)$ is defined by Eq.(2.2.1) for all $(\tau,\eta) \in \mathbb{C} \times \mathbb{R}^{n-1}$ with $\operatorname{Im}\tau \neq 0$, and it is a homogeneous function of degree $\prod_{k=1}^{m^+} d_k$.

2.2.3 COROLLARY. *Under Assumption 2.2.1 and the assumption that* $P$ *and* $B^k$ *are homogeneous, Eq.(2.1.1) is well-posed if and only if* $\Delta_B^+(\tau,\eta) \neq 0$ *for all* $(\tau,\eta) \in \mathbb{C} \times \mathbb{R}^{n-1}$, *with* $\operatorname{Im}\tau < 0$

PROOF: This follows from Thm.2.1.5 because $\Delta_B^+$ is a homogeneous function of $\tau,\eta$. □

2.2.4 DEFINITION. *The following notations will be used:*

$$\mathcal{A}_{\tau_0} = \{\ (\tau,\eta) \in \mathbb{C} \times \mathbb{R}^{n-1} \mid \operatorname{Im}\tau < -\tau_0\ \} .$$

$$\mathcal{A}_0 = \{\ (\tau,\eta) \in \mathbb{C} \times \mathbb{R}^{n-1} \mid \operatorname{Im}\tau < 0\ \} .$$

$$\bar{\mathcal{A}}_0 = \{\ (\tau,\eta) \in \mathbb{C} \times \mathbb{R}^{n-1} \mid \operatorname{Im}\tau \leq 0\ \} .$$

2.2.5 EXAMPLE.

We consider the equation

$$u_{tt} - u_{xx} - u_{yy} = 0 .$$

Then $P$ is given by $P(\tau,\xi,\eta) = -\tau^2 + \xi^2 + \eta^2$ . For any $(\tau,\eta) \in \mathcal{A}_0$ , $P = 0$ has two roots $\xi = \xi^+$ and $\xi^-$ with

$$\mathrm{Im}\,\xi^+ = -\mathrm{Im}\,\xi^- > 0 .$$

Thus $m^{\pm} = 1$ , and only one boundary condition needs to be prescribed at $x = 0$ . Assume that the boundary condition is given by

$$Bu = B(D_t, D_x, D_y)u = 0, \quad x = 0 ,$$

where $B$ is a homogeneous polynomial of $\tau, \xi, \eta$ . It is obvious that $\Delta_B^+(\tau,\eta) = B(\tau, \xi^+(\tau,\eta), \eta)$ . Therefore it follows from Cor.2.2.3 that the above problem is well-posed if and only if the equations

$$\tau^2 - \xi^2 - \eta^2 = 0 , \quad \text{and} \quad B(\tau,\xi,\eta) = 0$$

have no common solutions $(\tau,\xi,\eta) \in \mathbb{C} \times \mathbb{C} \times \mathbb{R}$ with $\mathrm{Im}\,\tau < 0$ and $\mathrm{Im}\,\xi > 0$ .

In general, the boundary operators $\{B^k\}$ in Eq.(2.1.1) are assumed to be a differential operator, i.e., $B^k$ is a polynomial of $(\tau,\xi,\eta)$ . However, if $B^k$ is an analytic function of $(\tau,\xi,\eta)$ , $\Delta_B^+$ in Eq.(2.2.1) is still well-defined and the previous discussion is also valid provided that $B^k(D_t, D_x, D_y)u$ is meaningful.

If $B = B(\tau,\xi,\eta)$ is some analytic function of $(\tau,\xi,\eta)$ , then a ***pseduo-differential operator*** $B(D_t, D_x, D_y)$ can be defined by [Nire73]

$$B(D_t, D_x, D_y)u = \int e^{iz\cdot\zeta} B(\zeta)\hat{u}(\zeta)d\zeta , \tag{2.2.3}$$

where $z = (t,x,y)$ , $\zeta = (\tau,\xi,\eta)$ , and $\hat{u}$ is the Fourier transform of $u$ . A difficulty arises when $B$ defined above is used as a boundary operator, because Eq.(2.2.3) requires information about $u$ away from the boundary $x = 0$ . On the other hand, as a boundary operator, $B$ need only be defined on the functions from the solution space of the equation $P(D_t, D_x, D_y)u = 0$ . Under Assumption 2.2.1, the solutions of $Pu = 0$ are made up of the Fourier modes $e^{i(\tau t+\xi x+\eta y)}$ . For such solutions, $B$ can be defined as

$$B(D_t, D_x, D_y)e^{i(\tau t+\xi x+\eta y)} = B(\tau, \xi, \eta)e^{i(\tau t+\xi x+\eta y)} .$$

In general we have

2.2.6 DEFINITION. *Assume*

$$u(t, x, y) = \sum_{j=1}^{m} \iint e^{i(\tau t+\xi_j(\tau,\eta)x+\eta y)}\hat{u}(\tau, x, \eta)d\sigma d\eta , \tag{2.2.4}$$

*where* $\hat{u}(\tau, x, \eta)$ *is the Fourier-Laplace transform of* $u$ *in* $t, y$ , *and* $\sigma = Re\,\tau$ . *Then a boundary operator* $B(D_t, D_x, D_y)$ *can be defined as*

$$B(D_t, D_x, D_y)u = \sum_{j=1}^{m} \iint e^{i(\tau t+\xi_j(\tau,\eta)x+\eta y)} B(\tau, \xi_j(\tau, \eta), \eta)\hat{u}(\tau, x, \eta)d\sigma d\eta . \tag{2.2.5}$$

$B(\tau, \xi, \eta)$ *is called the symbol of the operator* $B(D_t, D_x, D_y)$ .

Clearly the above definition is consistent with Eq.(2.2.3) if the *Dirac delta function* is used in the Fourier transform of Eq.(2.2.4) with respect to $x$ that appears in Eq.(2.2.3). Since no Fourier transform in $x$ is used in Eq.(2.2.5), it is a well-defined boundary operator. In the following, if a boundary operator is given by a non-polynomial function, its application to $u$ is implied by Def.2.2.6.

## 2.3 Absorption Properties of Boundary Operators

In this section, we investigate relations between the solution of a purely initial value problem and that of a mixed initial boundary value problem. Our interest is of course to approximate the solution of a purely initial value problem by that of an IBVP. We will prove a necessary and sufficient condition for the solutions of the initial value problem and the initial boundary value problem to be identical. The initial value problem is given by

$$P(D_t, D_x, D_y)u = 0\,, \qquad 0 < t, -\infty < x < +\infty, \quad y \in \mathbb{R}^{n-1} \qquad (2.3.1a)$$

$$\frac{\partial^k}{\partial t^k} u = f^k\,, \qquad t = 0, \quad k = 0, 1, \ldots, p-1\,, \qquad (2.3.1b)$$

where $f$ has compact support in $(0, +\infty)$ . In particular, it vanishes in a neighborhood of $x = 0$ .

If Eq.(2.3.1) is to be solved over the region $[0, +\infty)$ , some boundary condition is needed at $x = 0$ . Thus we consider

$$P(D_t, D_x, D_y)u = 0\,, \qquad 0 < t, x < +\infty, \quad y \in \mathbb{R}^{n-1} \qquad (2.1.1'a)$$

$$B^k(D_t, D_x, D_y)u = 0\,, \qquad x = 0, \quad k = 1, 2, \ldots, m^+ \qquad (2.1.1'b)$$

$$\frac{\partial^k}{\partial t^k} u = f^k\,, \qquad t = 0, \quad k = 0, 1, \ldots, p-1\,. \qquad (2.1.1'c)$$

2.3.1 DEFINITION. *We say* $\{B^k\}$, $k = 1, \ldots, m^+$ *is a well-posed boundary operator for Eq.(2.3.1) if the corresponding IBVP Eq.(2.1.1') is well-posed. We say* $\{B^k\}$ *is a perfectly absorbing boundary operator for Eq.(2.3.1) if the solutions of Eq.(2.3.1) and the corresponding Eq.(2.1.1') are identical.* $\{B^k\}$ *is said to be absorbing if the solution of Eq.(2.1.1') approximates that of Eq.(2.3.1).*

2.3.2 THEOREM. *Assume* $\{B^k\}$ *is a well-posed operator for Eq.(2.3.1). Then* $\{B^k\}$ *is a perfectly absorbing boundary operator for Eq.(2.3.1) if and only if* $\{B^k\}$ *annihilates the solution* $u$ *of Eq.(2.3.1), i.e.,* $B^k u = 0$ , $k = 1, \ldots, m^+$ .

PROOF: Suppose $\{B^k\}$ is a perfectly absorbing boundary operator. Then the solution $u$ of Eq.(2.3.1) is also a solution of Eq.(2.1.1'). In particular, $u$ satisfies the boundary condition Eq.(2.1.1'b), $B^k u = 0$, $k = 1, \ldots, m^+$.

On the other hand, if $\{B^k\}$ annihilates the solution $u$ of Eq.(2.3.1), then $u$ satisfies Eq.(2.1.1'). Thus it is the only solution of Eq.(2.1.1'). □

In the later study of solutions of Eq.(2.1.1'), the Fourier-Laplace transform will frequently be used. Since the nonzero Cauchy data $f$ in Eq.(2.1.1'c) is inconvenient for the Laplace transform in $t$, we reduce Eq.(2.1.1') to an equivalent problem where the effect of $f$ is absorbed into a forcing term in another boundary condition.

Since $f$ has compact support in $(0, +\infty)$, we can assume that there exists an $a > 0$ such that $\operatorname{supp} f \subset (a, +\infty)$. Then for any boundary operator $\{A^k\}, k = 1, \ldots, m^-$, the solution of Eq.(2.3.1) satisfies the following equation for some function $g$.

$$P(D_t, D_x, D_y)u = 0, \qquad 0 < t, 0 < x < a, \quad y \in \mathbb{R}^{n-1} \tag{2.3.2a}$$

$$A^k(D_t, D_x, D_y)u = g^k, \qquad x = a, \quad k = 1, 2, \ldots, m^- \tag{2.3.2b}$$

$$u \equiv 0, \qquad t \leq 0, \quad 0 \leq x \leq a. \tag{2.3.2c}$$

In fact, let $u$ be the solution of Eq.(2.3.1), and $g^k = A^k(D_t, D_x, D_y)u$, then $u$ satisfies Eq.(2.3.2). Thus, finding an absorbing boundary condition for Eq.(2.3.1) at $x = 0$ is equivalent to finding an absorbing boundary condition for Eq.(2.3.2), where the initial data is zero.

2.3.3 THEOREM. *Assume* $\{A^k\}$ $k = 1, \ldots, m^-$ *is a well-posed condition for the IBVP Eq.(2.3.2). If* $\{B^k\}$, *independent of* $g$, *is a perfectly absorbing boundary condition at* $x = 0$ *for Eq.(2.3.2), then* $\{B^k\}$ *is also a perfectly absorbing boundary condition for Eq.(2.3.1) at* $x = 0$.

**PROOF:** Since a solution of Eq.(2.3.1) is also a solution of Eq.(2.3.2) for some $g$, the fact that $\{B^k\}$ annihilates the solution of Eq.(2.3.2) implies that $\{B^k\}$ also annihilates the solution of Eq.(2.3.1). □

Due to Thm.2.3.3, we need only to study the problem of finding a perfectly absorbing boundary condition for Eq.(2.3.2). That is, we need to find a boundary operator $\{B^k\}$ such that the solution of

$$P(D_t, D_x, D_y)u = 0\,, \qquad 0 < t, 0 < x < a, \quad y \in \mathbb{R}^{n-1} \tag{2.3.3a}$$

$$A^k(D_t, D_x, D_y)u = g^k\,, \qquad x = a, \quad k = 1, 2, \dots, m^- \tag{2.3.3b}$$

$$B^k(D_t, D_x, D_y)u = 0\,, \qquad x = 0, \quad k = 1, 2, \dots, m^+ \tag{2.3.3c}$$

$$u \equiv 0\,, \qquad t \leq 0\,, \tag{2.3.3d}$$

and the solution of Eq.(2.3.2) are identical.

2.3.4 DEFINITION. *$\{B^k\}$ is well-posed for Eq.(2.3.3) if the IBVP Eq.(2.1.1') is well-posed. $\{B^k\}$ is said to be a perfectly absorbing boundary operator of Eq.(2.3.3) if the solutions of Eq.(2.3.3) and Eq.(2.3.2) are identical. $\{B^k\}$ is an absorbing boundary operator if the solution of Eq.(2.3.3) approximates that of Eq.(2.3.2).*

For now, we will not be concerned with the construction of a perfectly absorbing boundary condition. However, we shall give a necessary and sufficient condition for $\{B^k\}$ to be well-posed and perfectly absorbing.

2.3.5 THEOREM. *$\{B^k\}$ is a well-posed perfectly absorbing boundary operator* *for Eq.(2.3.3) if and only if there exists a* $\tau_0 > 0$ *, such that*

$$\Delta_B^+(\tau,\eta) \neq 0 \tag{2.3.4a}$$

$$N_B^-(\tau,\eta) = \mathcal{O}\,, \tag{2.3.4b}$$

*for all* $(\tau,\eta) \in \mathcal{A}_{\tau_0}$ .

In Thm.2.3.5, $\mathcal{O}$ is the $m^+ \times m^-$ zero matrix whose entries are all zero. $N_B^-$ is constructed in the same way as $N_B^+$ in § 2.1. That is, if $\{\phi_1^-, \ldots, \phi_{m^-}^-\}$ is a basis for the solution space $E^-(\tau,\eta)$ , then

$$N_B^-(\tau,\eta) = B[\phi_1^-(0), \ldots, \phi_{m^-}^-(0)]\,.$$

PROOF: We first assume that $\{B^k\}$ is well-posed and perfectly absorbing. Then Eq.(2.3.4a) follows directly from Thm.2.1.5. Since $\{B^k\}$ is also perfectly absorbing, by Thm.2.3.2, it annihilates the solution of Eq.(2.3.2). The solution $u$ of Eq.(2.3.2) can be found explicitly, and its Fourier-Laplace transform is given by

$$\hat{u}(\tau,x,\eta) = \sum_{j=1}^{m^-} c_j^- \phi_j^- \,, \tag{2.3.5}$$

where $\phi_j^- \in E^-$ and $(\tau,\eta) \in \mathcal{A}_{\tau_0}$ . Therefore $B^k u = 0$ implies

$$N_B^-(\tau,\eta)c^- = 0 \tag{2.3.6}$$

for all $c^-$ . Thus the entries of $N_B^-$ must be all zero.

The converse can be proved by employing Thm.2.1.5 and Thm.2.3.2 once more. □

If we denote a norm of a matrix by $\|\cdot\|$ , then Thm.2.3.5 shows that perfect absorption is equivalent to $\|N_B^-(\tau,\eta)\| = 0$ for all $(\tau,\eta) \in \mathcal{A}_{\tau_0}$ . A natural question to be asked is whether the solution of Eq.(2.3.3) approximates that of

Eq.(2.3.2) if $\|N_B^-(\tau,\eta)\|$ is nonzero but very small. To answer this question, we consider a sequence of boundary conditions $\{B_d^k\}$ ; we hope to show that if $\|N_{B_d}^-\| \to 0$ , then $\{B_d^k\}$ approximates a perfectly absorbing boundary operator very well as $d$ becomes very large.

Let $B_d^k = Q_d B^k$ , $k = 1, \ldots, m^+$, $d = 1, 2, \ldots$ , where $Q_d = Q_d(\tau,\xi,\eta)$ is some homogeneous polynomial of $(\tau,\xi,\eta)$ . Then from the definition of $N_{B_d}^{\pm}$ , it is easy to see that under Assumption 2.2.1

$$N_{B_d}^{\pm}(\tau,\eta) = N_B^{\pm}(\tau,\eta)\mathbb{Q}_d^{\pm}(\tau,\eta),$$

where $\mathbb{Q}_d^{\pm}(\tau,\eta)$ is an $m^{\pm} \times m^{\pm}$ diagonal matrix whose diagonal elements are $Q_d(\tau,\xi_j^{\pm}(\tau,\eta),\eta)$ , $j = 1, \ldots, m^{\pm}$ . We also assume that $P$ and $B^k$ are homogeneous polynomials of $(\tau,\xi,\eta)$ .

2.3.6 THEOREM. *Suppose that $\{A^k\}$, $\{B_d^k\}$ are well-posed, and that there exists $\delta > 0$ such that*

$$|\Delta_A^-(\tau,\eta)| \geq \delta, \quad |\Delta_{B_d}^+(\tau,\eta)| \geq \delta, \tag{2.3.7}$$

*for all $(\tau,\eta) \in \mathcal{A}_0$ with $|\tau|^2 + |\eta|^2 = 1$ . Also assume that $\|\mathbb{Q}_d^-\| \, \|(\mathbb{Q}_d^+)^{-1}\| \leq \varepsilon_d$ for all $(\tau,\eta) \in \mathcal{A}_0$ where $\varepsilon_d$ is independent of $(\tau,\eta)$ and $\varepsilon_d \to 0$ as $d \to \infty$ . Then there exists a $d_0$ such that the problem Eq.(2.3.3) (with $\{B^k\}$ replaced by $\{B_d^k\}$ ) has a unique solution $u_d$ for all $d \geq d_0$ . Furthermore, for any $T > 0$*

$$u_d \to u \qquad \text{in} \quad \mathcal{L}^2_{[0,T]\times[0,a]\times\mathbb{R}^{n-1}}, \quad \text{as } d \to +\infty,$$

*where $u$ is the solution of Eq.(2.3.2).*

PROOF: In Eq.(2.3.3) (with $\{B^k\}$ replaced by $\{B_d^k\}$ ), take Laplace transform in $t$ and Fourier transform in $y$ . Then the solution of the transformed equation is given by

$$\hat{u}_d = \sum c_j^+ \phi_j^+ + \sum c_j^- \phi_j^- \,, \tag{2.3.8}$$

subject to the boundary conditions

$$\begin{aligned} A^k(\tau, D_x, \eta)\hat{u}_d &= \hat{g}^k, \qquad \text{at} \quad x = a \\ B_d^k(\tau, D_x, \eta)\hat{u}_d &= 0, \qquad \text{at} \quad x = 0\,. \end{aligned} \tag{2.3.9}$$

When Eq.(2.3.8) is substituted into Eq.(2.3.9), the result is

$$N_A^+ c^+ + N_A^- c^- = \hat{g} \tag{2.3.10a}$$

$$N_{B_d}^+ c^+ + N_{B_d}^- c^- = 0\,, \tag{2.3.10b}$$

where $c^\pm = (c_1^\pm, \ldots, c_{m^\pm}^\pm)$ .

By Thm.2.1.5, $N_{B_d}^+$ is invertible. Solving for $c^+$ from Eq.(2.3.10b) and inserting it into Eq.(2.3.10a), we obtain

$$\{-N_A^+ (N_{B_d}^+)^{-1} N_{B_d}^- + N_A^-\}\, c^- = \hat{g}$$

or

$$\{I - (N_A^-)^{-1} N_A^+ (N_{B_d}^+)^{-1} N_{B_d}^-\}\, c^- = (N_A^-)^{-1} \hat{g}\,. \tag{2.3.11}$$

In Eq.(2.3.11), $N_A^-$ is invertible because $\{A^k\}$ is well-posed.

Let us define

$$N = (N_A^-)^{-1} N_A^+ (N_{B_d}^+)^{-1} N_{B_d}^- \,.$$

Then

$$\|N\| \le \|(N_A^-)^{-1}\| \|N_A^+\| \|(N_B^+)^{-1}\| \|N_B^-\| \|(\mathbb{Q}_d^+)^{-1}\| \|\mathbb{Q}_d^-\| \,.$$

We shall show that there exists some $M > 0$ such that

$$\|N\| \le M\varepsilon_d \,, \quad \text{for all } (\tau, \eta) \in \mathcal{A}_0 \,.$$

The above estimate holds if $(\tau, \eta)$ is in some compact set, i.e., if $|\tau|^2 + |\eta|^2$ is bounded, because $N$ is continuous in $(\tau, \eta)$. We shall show the above estimate also holds when $|\mathrm{Re}\,\tau|^2 + |\eta|^2 \to \infty$.

From their construction, we see that each entry of $N_B^{\pm}$ is some polynomial in $(\tau, \xi, \eta)$, and that of $N_A^{\pm}$ is some polynomial times an exponential term $e^{i\xi_j^{\pm} a}$. Thus $\|(N_B^+)^{-1}\| \|N_B^-\|$ can be bounded by a rational function of $(\tau, \xi, \eta)$ and $\|(N_A^-)^{-1}\| \|N_A^+\|$ can be bounded by a rational function multiplied by $e^{i(\xi_0^+ - \xi_0^-)a}$, where $\mathrm{Im}\,\xi_0^+ > 0$, $\mathrm{Im}\,\xi_0^- < 0$ and $a > 0$. Because $r(\tau, \xi, \eta) e^{i(\xi_0^+ - \xi_0^-)a}$ decays to zero for large $|\mathrm{Re}\,\tau|^2 + |\eta|^2$ for any rational $r$ whose denominator does not vanish, and because Eq.(2.3.7) holds, we conclude that

$$\|N\| \le M\varepsilon_d \tag{2.3.12}$$

for all $(\tau, \eta) \in \mathcal{A}_0$, where $M > 0$ is independent of $(\tau, \eta)$ and $d$.

Now let $d$ be large enough so that $\|N\| < 1$. Then $c^-$ can be obtained from solving Eq.(2.3.11):

$$c^- = \sum_{h=0}^{\infty} N^h (N_A^-)^{-1} \hat{g}\,.$$

In turn $c^+$ can be found as

$$c^+ = (N_{B_d}^+)^{-1} (N_{B_d}^-) \sum_{h=0}^{\infty} N^h (N_A^-)^{-1} \hat{g}\,.$$

Inserting these into Eq.(2.3.8), we find that

$$\hat{u}_d = \hat{u} + \sum_{h=1}^{\infty} \hat{v}_d^h\,, \tag{2.3.13}$$

where $\hat{u} = \sum c_j^- \phi_j^- = \phi^- \cdot c^- = \phi^- (N_A^-)^{-1} \hat{g}$ and $\phi^- = [\phi_1^-, \dots, \phi_{m^-}^-]$. Similarly, $\hat{v}_d^h$ is given by

$$\hat{v}_d^h = \phi^- N^h (N_A^-)^{-1} \hat{g} + \phi^+ (N_{B_d}^+)^{-1} N_{B_d}^- N^{h-1} (N_A^-)^{-1} \hat{g}\,.$$

We claim that $\hat{u}_d$ given above is independent of the choice of the basis $\phi^-$, $\phi^+$ . In fact, if $\tilde{\phi}^-$, $\tilde{\phi}^+$ is another basis, then there is some nonsingular matrix $W^\pm$ such that

$$\tilde{\phi}^\pm = \phi^\pm W^\pm .$$

Then

$$\tilde{N}_A^\pm = N_A^\pm W^\pm, \quad \tilde{N}_{B_d}^\pm = N_{B_d}^\pm W^\pm,$$

and

$$\begin{aligned}\tilde{N} &= (\tilde{N}_A^-)^{-1}(\tilde{N}_A^+)(\tilde{N}_{B_d}^+)^{-1}(\tilde{N}_{B_d}^-)\\ &= (W^-)^{-1}(N_A^-)^{-1}W^+(N_A^+)(W^+)^{-1}(N_{B_d}^+)^{-1}(N_{B_d}^-)W^-\\ &= (W^-)^{-1}NW^- .\end{aligned}$$

Therefore

$$\tilde{\hat{u}} = \tilde{\phi}^-(\tilde{N}_A^-)^{-1}\hat{g} = \phi^- W^-(W^-)^{-1}N_A^-\hat{g} = \hat{u} .$$

Similarly, $\tilde{\hat{v}}_d^h = \hat{v}_d^h$ . This shows that $\hat{u}_d$ is independent of the choice of the basis.

The inverse transform of Eq.(2.3.13) exists and will be the solution of Eq.(2.3.3) if we can show that $\hat{u}_d$ is a holomorphic function of $(\tau, \eta) \in \mathcal{A}_0$ . In fact, each element of $N_A^\pm$, $N_{B_d}^\pm$ is a polynomial or an exponential function of $(\tau, \xi, \eta)$ , and therefore an entire function of $(\tau, \xi, \eta)$ . This is also true for the inverse matrices if $(\tau, \eta) \in \mathcal{A}_0$ . Therefore, $\hat{u}_d$ is a holomorphic function of $(\tau, \xi, \eta)$ in the domain of definition. But $\xi = \xi(\tau, \eta)$ is a root of the equation $P(\tau, \xi, \eta) = 0$ , and is therefore singular at the branch points. This means that $\hat{u}_d$ is possibly singular at the branch points of the roots. These singular points are isolated. In a neighborhood of a singular point, i.e., a branch point, the root $\xi(\tau, \eta)$ remains finite. As a consequence, $\hat{u}_d$ remains finite. However, if a function is finite in a neighborhood of an isolated singular point, the singular point is removable. This shows that there are at worst isolated removable singular points of $\hat{u}_d$ . Redefining values of the function at these points if necessary, we have shown that $\hat{u}_d$ is a holomorphic function. Therefore its inverse Fourier-Laplace transform, denoted by $u_d$ , exists, and is the solution of Eq.(2.3.3). This argument resembles that of Hersh [Hers64].

It remains to show that $u_d \to u$ . It is easy to show that $\hat{u}$ is the Fourier-Laplace transform of the solution $u$ of Eq.(2.3.2), and

$$u_d - u = \sum_{h=0}^{\infty} v_d^h ,$$

where $v_d^h$ is the inverse Fourier-Laplace transform of $\hat{v}_d^h$ . Applying the Paley-Wiener theorem [Pete83] in a standard fashion, we see that for any $T > 0$ , there exists an $H = H(T)$ such that $v_d^h = 0$ for $t < T$ if $h > H$ . Thus for $(t, x, y) \in [0, T] \times [0, a] \times \mathbb{R}^{n-1}$ ,

$$u_d - u = \sum_{h=0}^{H} v_d^h .$$

Therefore

$$\begin{aligned} \|u_d - u\|_{\mathcal{L}^2_{[0,T]\times[0,a]\times\mathbb{R}^{n-1}}} &\leq \sum_{h=0}^{H} \|v_d^h\|_{\mathcal{L}^2_{[0,T]\times[0,a]\times\mathbb{R}^{n-1}}} \\ &\leq \sum_{h=0}^{H} \|v_d^h\|_{\mathcal{L}^2_{[0,+\infty)\times[0,a]\times\mathbb{R}^{n-1}}} . \end{aligned}$$

Since

$$\|v_d^h\|^2_{\mathcal{L}^2_{[0,+\infty)\times[0,a]\times\mathbb{R}^{n-1}}} = \|\hat{v}_d^h\|^2_{\mathcal{L}^2_{[0,+\infty)\times[0,a]\times\mathbb{R}^{n-1}}} ,$$

by Parseval's equality, we only need to show that

$$\|\hat{v}_d^h\|^2_{\mathcal{L}^2_{[0,+\infty)\times[0,a]\times\mathbb{R}^{n-1}}} \to 0 , \qquad \text{as} \quad d \to \infty .$$

Using the same argument that was used to show Eq.(2.3.12), we can show that

$$|\hat{v}_d^h| \leq M_h' \varepsilon_d |\hat{g}|,$$

where $M_h' > 0$ is independent of $(\tau, \eta)$ and $d$. Thus

$$\|\hat{v}_d^h\|^2_{\mathcal{L}^2_{[0,+\infty)\times[0,a]\times\mathbb{R}^{n-1}}} \leq a M_h' \varepsilon_d \|\hat{g}\|^2_{\mathcal{L}^2_{[0,+\infty)\times\mathbb{R}^{n-1}}} \to 0$$

as $d \to +\infty$. $\square$

As we shall see in a later chapter, the value $\|Q_d^-\| \|(Q_d^+)^{-1}\|$ is called the absorbing factor. The importance of Thm.2.3.6 is that if the absorbing factor can be made smaller and smaller, the solution of Eq.(2.3.3) better and better approximates that of Eq.(2.3.2).

The condition Eq.(2.3.7) is referred to as the Uniform Kreiss Condition. This condition implies the well-posedness of the IBVP as given by Thm.2.1.5, but the former is stronger. A detailed discussion of the Uniform Kreiss Condition will be given in the next chapters.

In Thm.2.3.6, $\{B_d^k\}$ does not have to be in the stated form. What is essential is the condition that $\|N\|$ can be made arbitarily small uniformly with respect to $(\tau, \eta) \in \mathcal{A}_0$.

# CHAPTER 3

# WAVE PROPAGATION IN INITIAL BOUNDARY VALUE PROBLEMS

Hyperbolic equations can be studied from a wave propagation point of view. By doing so, many mathematical terms used in the previous chapter, for example, the well-posedness condition, have physical interpretations. In this chapter, we will connect the well-posedness and absorption properties of a boundary condition to the theory of wave propagation. In section one, we introduce the concept of group velocity. Section two deals with the classification of leftgoing and rightgoing solutions. In section three, we discuss the Uniform Kreiss Condition. Finally, in section four, we will derive the reflection coefficient of a boundary condition, which provides a natural quantitative indication of absorption properties of a boundary condition.

In the remaining part of the thesis, unless otherwise stated, we will always assume that the equation in Eq.(2.1.1a) is homogeneous, i.e., $P$ is a homogeneous polynomial of $(\tau, \xi, \eta)$ .

## 3.1 Group Velocity

In the previous chapter, we studied solutions by using the Fourier-Laplace transform. These solutions are composed of the Fourier modes

$$e^{i(\tau t+\xi x+\eta y)}, \tag{3.1.1}$$

where $(\tau, \xi, \eta) \in \mathbb{C} \times \mathbb{C} \times \mathbb{R}^{n-1}$ , $\operatorname{Im} \tau < 0$ . Here we study the behavior of the mode Eq.(3.1.1) in the limit as $\operatorname{Im} \tau \to 0$ . For $(\tau, \xi, \eta) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{n-1}$ , inserting Eq.(3.1.1) into Eq.(2.1.1a) results in

$$P(\tau,\xi,\eta)=0\,. \tag{3.1.2}$$

Eq.(3.1.2) is referred to as the *dispersion relation* of Eq.(2.1.1a).

The form Eq.(3.1.1) is naturally related to the propagation of wave solutions of Eq.(2.1.1a). For real $(\tau,\xi,\eta)$ , Eq.(3.1.1) represents a plane wave solution, its frequencies and wave numbers are governed by the dispersion relation Eq.(3.1.2).

3.1.1 DEFINITION.

Define the characteristic variety [Higd86b] $\mathcal{V}$ by

$$\mathcal{V}=\left\{(\tau,\xi,\eta)\in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^{n-1}\ \middle|\ P(\tau,\xi,\eta)=0\ \right\}.$$

Its projection on the $(\tau,\eta)$ space is denoted by $\Gamma$ :

$$\Gamma=\left\{(\tau,\eta)\in \mathbb{R}\times\mathbb{R}^{n-1}\ \middle|\ (\tau,\xi,\eta)\in\mathcal{V}\ \right\}.$$

For each $(\tau,\xi,\eta)\in\mathcal{V}$, Eq.(3.1.1) is a wave-like solution; it is constant along the lines $(t,x,y)\cdot(\tau,\xi,\eta)=\text{constant}$ . This means that the wave propagates at the velocity

$$c=\frac{-\tau}{\xi^2+|\eta|^2}(\xi,\eta)\,.$$

This is the *phase velocity*, at which the crests and troughs of waves propagate. When waves of various wave lengths are superimposed, the energy of a wave train propagates at a different velocity; namely, the *group velocity* given by

$$C=-\nabla\,\tau=(-\frac{\partial\tau}{\partial\xi},-\nabla_\eta\,\tau)\,. \tag{3.1.3}$$

In order to motivate this phenomenon, we consider a signal of superimposition of two sinusoidal waves in one dimension. Suppose the signal is

$$u = \sin(\tau_1 t + \xi_1 x) + \sin(\tau_2 t + \xi_2 x)\,,$$

where $\tau_1 \approx \tau_2$ , $\xi_1 \approx \xi_2$ . Then the composite wave

$$u = 2\sin\frac{\xi_1+\xi_2}{2}\Big(\frac{\tau_1+\tau_2}{\xi_1+\xi_2}t + x\Big)\sin\frac{\xi_1-\xi_2}{2}\Big(\frac{\tau_1-\tau_2}{\xi_1-\xi_2}t + x\Big)$$

is a single sinusoidal wave with wave number $\frac{\xi_1+\xi_2}{2}$ multipled by a slowly varying sinusoidal envelope with wave number $\frac{\xi_1-\xi_2}{2}$ , as illustrated in Fig.3.1.1. It is evident that the envelope which contains the energy is moving at the speed

$$-\frac{\tau_1-\tau_2}{\xi_1-\xi_2}.$$

In the limit $\xi_1 \to \xi_2$, $\tau_1 \to \tau_2$ , the above quantity approaches $-\frac{\partial\tau}{\partial\xi}$ ; the $x$ - component of Eq.(3.1.3). For more discussion of group velocity, see [Whit74] and [Ligh78].

Related to the group velocity is the theory of propagation of singularities ([Nire73] and [Tayl81]), which is a discussion of the propagation of the high-frequency portions of the solution. The singularities of a solution propagate along *bicharacteristic curves*. Let $z = (t,x,y)$ and $\zeta = (\tau,\xi,\eta)$ . Then the bicharacteristic curves are the curves $(z(s),\zeta(s))$ that satisfy, in the case of constant coefficients,

$$\begin{aligned}\dot z &= \nabla_\zeta P(\zeta)\\ \dot\zeta &= \quad 0\,,\end{aligned}$$

subject to $P(\zeta) = 0$ .

The path of the propagation of singularities in $(t,x,y)$ space is $z(\zeta)$ ; its

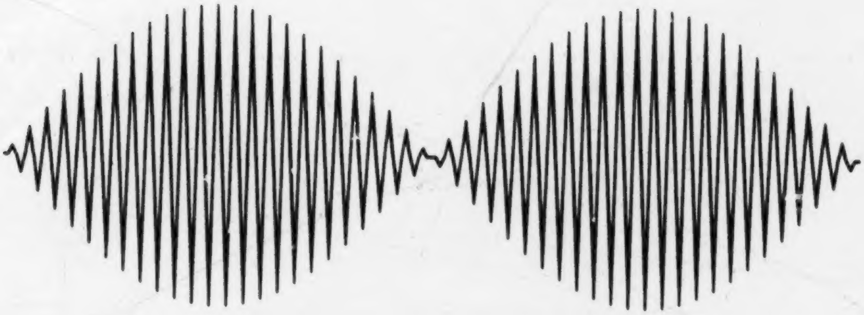

Figure 3.1.1 Group Velocity

direction is parallel to $\nabla_\zeta P(\zeta)$ . The projection of the latter onto the $(x, y)$ - plane is $(P_\xi, \nabla_\eta P)$ which is parallel to the group velocity.

The above discussion is still valid in the case of variable coefficients, provided that they are *slowly varying.* In the case of variable coefficients, wave propagations are governed by the *ray theory* as discussed in [Whit74] and [Ligh78].

## 3.2 Leftgoing and Rightgoing Solutions

In chapter 2, we have studied intensively the Fourier modes

$$e^{i(\tau t+\xi_j^+ x+\eta y)}, \quad e^{i(\tau t+\xi_j^- x+\eta y)},$$

where $\operatorname{Im}\xi_j^+ > 0$, $\operatorname{Im}\xi_j^- < 0$ , for $(\tau, \eta) \in \mathcal{A}_0$ . Now we consider the limiting case as $\operatorname{Im}\tau \to 0$ . In the following discussion, the index $j$ is irrelevant, so it will be dropped from the notations.

In the case when $\operatorname{Im}\tau < 0$, the group velocity of a wave packet $e^{i(\tau t+\xi x+\eta y)}$ is not defined. However, there is a *speed of translation of amplitude* naturally associated with such modes. Let $\Delta t$ be an increment in $t$, and $\Delta x$ be the corresponding increment in $x$ such that the amplitude of the plane wave

$$|e^{i(\tau t+\xi x+\eta y)}| \tag{3.2.1}$$

is kept constant. Then we define the quantity

$$\overset{\circ}{C}_x = \frac{\Delta x}{\Delta t}$$

to be the speed of translation of amplitude in the $x$-direction. That is, the amplitude Eq.(3.2.1) is constant along the lines $x - \overset{\circ}{C}_x t = \text{constant}$. Therefore, solving for $\dfrac{\Delta x}{\Delta t}$ from

$$|e^{i(\tau t+\xi x+\eta y)}| = |e^{i(\tau(t+\Delta t)+\xi(x+\Delta x)+\eta y)}|,$$

we obtain

$$\overset{\circ}{C}_x(\tau,\xi,\eta) = -\frac{\operatorname{Im}\tau}{\operatorname{Im}\xi}.$$

With this definition, a wave is naturally called a *rightgoing wave* if $\overset{\circ}{C}_x \geq 0$; it is called a *leftgoing wave* if $\overset{\circ}{C}_x \leq 0$.

We now relate the concept of leftgoing or rightgoing waves to the classification of the roots of the dispersion relation Eq.(3.1.2) as given in Eq.(2.1.5). For each $(\tau,\eta) \in \mathcal{A}_0$, the roots of Eq.(3.1.2) are given by

$$\xi_j^{\pm}, \quad j = 1,\ldots, m^{\pm},$$

where $\operatorname{Im}\xi_j^+ > 0$ and $\operatorname{Im}\xi_j^- < 0$. Consequently, the associated speeds of translation satisfy

$$\mathring{C}_x(\tau,\xi^+,\eta) > 0\,, \quad \mathring{C}_x(\tau,\xi^-,\eta) < 0\,.$$

Thus the mode $(\tau,\xi^+,\eta)$ is rightgoing while $(\tau,\xi^-,\eta)$ is leftgoing.

As $\operatorname{Im}\tau \to 0$, we obtain either $\operatorname{Im}\xi^+ > 0$ or $\operatorname{Im}\xi^+ = 0$. In the former case, $\xi^+(\tau,\eta)$ is complex, and $\mathring{C}_x(\tau,\xi^+,\eta) = 0$ when $(\tau,\eta) \in \mathbb{R} \times \mathbb{R}^{n-1}$. Such modes are called ***evanescent waves*** because the amplitude Eq.(3.2.1) decays exponentially as $x \to +\infty$.

We now consider the more delicate case where $\operatorname{Im}\xi^+(\tau,\eta) \to 0$ as $\operatorname{Im}\tau \to 0$. Thus we assume that there is $(\tau_0,\xi_0^+) \in \mathbb{R}\times\mathbb{R}$, such that $\xi^+(\tau,\eta) \to \xi_0^+$ as $\tau \to \tau_0$. It is clear from the definition that $(\tau_0,\xi_0^+,\eta) \in \mathcal{V}$, the characteristic variety. Therefore there is a group velocity associated with it. Such a mode is referred to as a ***travelling wave*** because its amplitude is conserved in the propagation process. Since the mode $(\tau_0,\xi_0^+,\eta)$ is the limit of a rightgoing mode, we expect it to be rightgoing as well, i.e., it should have a nonnegative group velocity in the $x$-direction. We will show this is indeed the case.

To study the limit of $\mathring{C}_x$, we need to assume the analytic dependence of $\tau$ and $\xi^\pm$. More precisely, we assume that there is a neighborhood of $\tau_0$ such that $\xi^\pm$ and $\tau$ are analytic functions of each other in the intersection of this neighborhood with the lower-half complex plane $\operatorname{Im}\tau < 0$.

3.2.1 THEOREM. *Let* $\tau$ *and* $\xi^\pm$ *have the analytic dependence as stated above. If* $C_x(\tau_0,\xi_0^\pm,\eta)$ *is the group velocity associated with* $(\tau_0,\xi_0^\pm,\eta)$, *then*

$$\lim_{\tau\to\tau_0} \mathring{C}_x(\tau,\xi^\pm,\eta) = C_x(\tau_0,\xi_0^\pm,\eta)\,.$$

*Consequently,* $C_x(\tau_0,\xi_0^+,\eta) \geq 0$ *(* $C_x(\tau_0,\xi_0^-,\eta) \leq 0$ *).*

**Proof:** By assumption, there is a neighborhood of $\tau_0$ where $\tau$ and $\xi^+$ are analytic functions of each other. If $\xi^+ = \xi_0^+ + \Delta\xi$ , then

$$\tau = \tau(\xi_0^+ + \Delta\xi) = \tau_0 + \frac{\partial\tau}{\partial\xi}\bigg|_{(\tau_0,\xi_0^+,\eta)} \cdot \Delta\xi + o(\Delta\xi)\,.$$

Since $C_x(\tau_0,\xi_0^+,\eta) = -\frac{\partial\tau}{\partial\xi}\Big|_{(\tau_0,\xi_0^+,\eta)}$ is real, we have

$$\operatorname{Im}\tau = -C_x(\tau_0,\xi_0^+,\eta)\operatorname{Im}\xi^+ + o(\Delta\xi)\,.$$

In obtaining the above, we have used $\operatorname{Im}\xi^+ = \operatorname{Im}\Delta\xi$ . The above equation leads to

$$\overset{\circ}{C}_x(\tau,\xi^+,\eta) = -\frac{\operatorname{Im}\tau}{\operatorname{Im}\xi^+} = C_x(\tau_0,\xi_0^+,\eta) + o(\Delta\xi)\,,$$

and the theorem follows. □

Trefethen [Tref82] has proved a similar result for finite difference schemes resulting from the discretization of hyperbolic systems.

3.2.2 Definition. *Extend the definition of* $\overset{\circ}{C}_x$ *to the case* $\operatorname{Im}\tau = 0$ *by Thm.3.2.1. Then for any* $(\tau,\eta) \in \bar{\mathcal{A}}_0$ *, the plane wave solution* $e^{i(\tau t+\xi x+\eta y)}$ *is rightgoing (leftgoing, resp.), if* $\overset{\circ}{C}_x(\tau,\xi,\eta) \geq 0$ *(* $\overset{\circ}{C}_x(\tau,\xi,\eta) \leq 0$ *, resp.). Its speed of propagation in the* $x$ *-direction is* $\overset{\circ}{C}_x(\tau,\xi,\eta)$ .

## 3.3 Strongly Well-Posedness

In chapter 2 we have seen that the IBVP Eq.(2.1.1) is well-posed (in the sense of Hersh, Thm.2.1.5) if and only if

$$\Delta_B^+(\tau,\eta) \neq 0\,, \quad \text{for all} \quad (\tau,\eta) \in \mathcal{A}_0\,. \tag{3.3.1}$$

If this condition holds, the solution can be estimated in terms of the prescribed data in the form of Eq.(2.1.2) (see [Duff76], [Hers64] and [Saka82]). It seems that this result should be quite satisfactory, but as the following example shows, it is not quite so.

### 3.3.1 EXAMPLE.

Consider the first-order hyperbolic equation

$$\begin{aligned} u_t + u_x &= 0\,, \quad t, x \geq 0 \\ Bu = u_x &= 0\,, \quad x = 0 \\ u &= f\,, \quad t = 0\,. \end{aligned} \tag{3.3.2}$$

It is easy to show that $\Delta_B^+(\tau) = -\tau \neq 0$, for all $\tau \in \mathbb{C}$ such that $\operatorname{Im}\tau < 0$. Thus according to Thm.2.1.5, the problem Eq.(3.3.2) is well-posed.

Suppose $f$ has values as shown in Fig.3.3.1(a), i.e.,

$$f(x) = \begin{cases} 1\,, & \text{for } 0 < x < \varepsilon_1 \\ 0\,, & \text{for } x > \varepsilon_2 \\ \text{smooth}, & \text{for } x \in [\varepsilon_1, \varepsilon_2], \end{cases}$$

where $0 < \varepsilon_1 < \varepsilon_2$ are small. Let us study the behavior of the solution as $t$ increases. At the boundary $x = 0$, the solution $u = 1$ satisfies the boundary condition $u_x = 0$, therefore $u = 1$ persists. At the other end, the wave front will travel at the group velocity which is 1 in this case. After some time, say $t = 1$, the solution has a shape shown in Fig.3.3.1(b). If we denote by $\|\cdot\|$ the $\mathcal{L}^2_{[0,+\infty)}$ norm, then the solution satisfies

$$\|u(t)\|^2 \geq t + \varepsilon_1\,, \quad \|f\|^2 \leq \varepsilon_2\,. \tag{3.3.3}$$

Thus an estimate of the form

$$\|u(t)\| \leq C(t)\|f\| \tag{3.3.4}$$

is *never* possible if $\varepsilon_1 \approx \varepsilon_2$ . This may lead one to question the validity of Thm.2.1.5. However, the problem Eq.(3.3.2) is indeed well-posed, because one easily verifies that

$$\|u_x(t)\| \leq \|f_x\|, \quad \text{for all } t \geq 0. \tag{3.3.5}$$

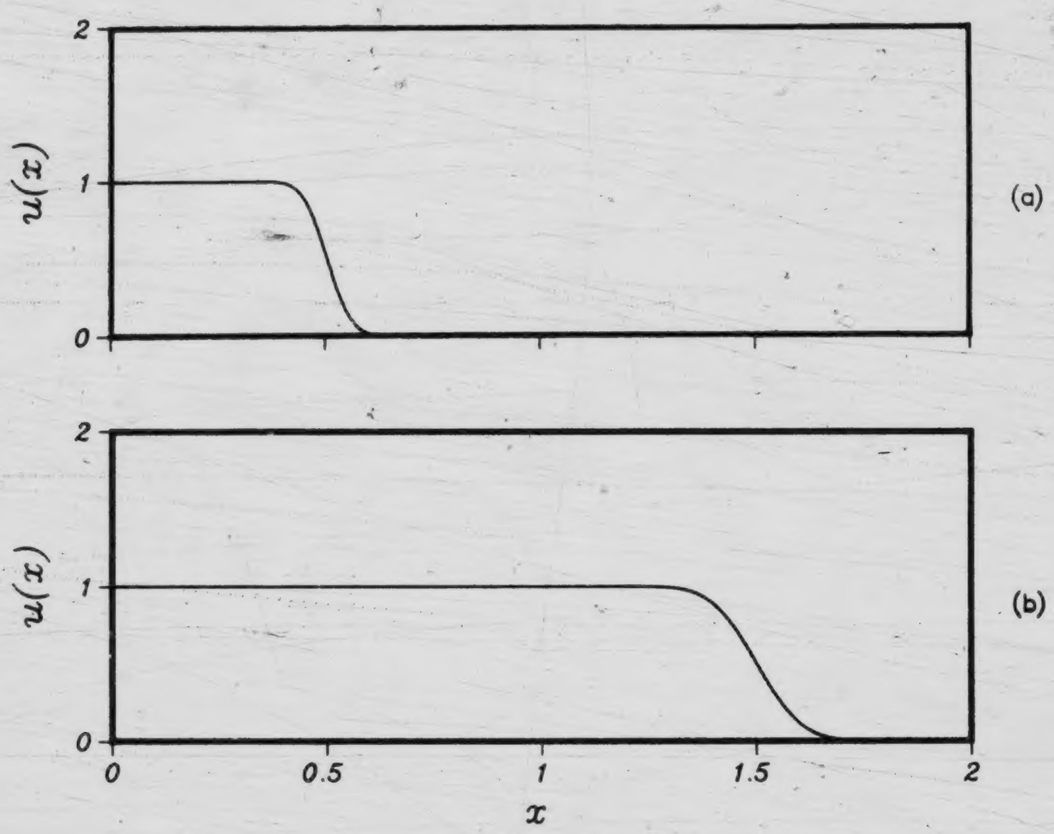

**Figure 3.3.1** Weak ill-posedness
(a) $t = 0$ , (b) $t = 1$ .

This example shows that although a problem is well-posed, in the sense of Hersh, one cannot expect a strong estimate of the form Eq.(3.3.4) where the solution can be estimated in terms of the prescribed data only (not their derivatives).

In general, one has to settle for the weaker form of Eq.(3.3.5) where the estimate is given in terms of their derivatives as well as the initial data. Note that in the definition of well-posedness in chapter 2, we have deliberately left the norm unspecified.

To answer the question why the strong estimate Eq.(3.3.4) is not possible, we note that $\Delta_B^+(\tau) = 0$ admits a root $\tau_0 \in \mathbb{R}$, namely, $\tau_0 = 0$. This root corresponds to a travelling mode with the group velocity $1$. Thus Eq.(3.3.2) admits a rightgoing wave which carries energy into the space domain $[0, +\infty)$, making Eq.(3.3.4) impossible.

The essence of the above example applies to more general problems. Recall from the proof of Thm.2.3.6, that the solution of Eq.(2.3.3) is given by

$$\hat{u} = N\hat{g}\,, \tag{3.3.6}$$

where $N$ has $(N_B^+)^{-1}$ as a factor. The condition Eq.(3.3.1) insures invertibility of $N_B^+$ for $\operatorname{Im}\tau < 0$. However, as $|\operatorname{Re}\tau|^2 + |\eta|^2 \to +\infty$ (or equivalently, $\operatorname{Im}\tau \to 0$ by homogeneity of $N_B^+$), $N_B^+(\tau, \eta)$ may become more and more singular, causing some components of $\hat{u}$ to become very large for high frequencies. Thus estimating $\hat{u}$ in terms of $g$ may not be possible. Since the growth in $(N_B^+)^{-1}$ can at worst be the inverse of a polynomial in $(\tau, \eta)$, and multiplying $\hat{g}$ by $\tau$, $\eta$ corresponds to differentiating $g$ with respect to $t$, $y$, the large components of $\hat{u}$ can be estimated in terms of the derivatives of $g$. However, whether this situation actually occurs depends on the term $\hat{g}$. If it does occur, we say there is a loss of derivatives at the boundary, or weak ill-posedness.

To prevent weak ill-posedness, we need a stronger condition than Eq.(3.3.1), namely, $\Delta_B^+(\tau, \eta) \neq 0$ for all $(\tau, \eta) \in \bar{\mathcal{A}}_0$, the closure of $\mathcal{A}_0$. This is the *Uniform Kreiss Condition (UKC)* as we have already seen in Thm.2.3.6, Eq.(2.3.7). If the UKC holds, then the matrix $N$ in Eq.(3.3.6) is uniformly bounded for all $(\tau,\eta) \in \mathcal{A}_0$ . Thus obtaining an estimate in terms of $g$ is possible.

3.3.2 DEFINITION. *IBVP Eq.(2.1.1) is said to be strongly well-posed, or well-posed in the sense of Kreiss, if*

$$\Delta_B^+(\tau,\eta) \neq 0\,, \quad \textit{for all} \quad (\tau,\eta) \in \bar{\mathcal{A}}_0\,.$$

Kreiss [Krei70] studied first-order hyperbolic systems with variable coefficients. He showed that the UKC is a necessary and sufficient condition for well-posed mixed initial boundary value problems. The criterion given in Def.3.3.2 is often referred to as *normal mode analysis.* A detailed discussion of Kreiss' theory in terms of wave propagation can be found in [Higd86b]; a similar treatment for stability of finite difference schemes can be found in [Tref82].

3.3.3 DEFINITION. $(\tau,\xi,\eta) \in \mathcal{V} \subset \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{n-1}$ *is said to be a generalized eigenvalue of Eq.(2.1.1) if*

$$\Delta_B^+(\tau,\eta) = 0\,.$$

Some eigenvalues and generalized eigenvalues deserve a special attention. They are those $(\tau,\xi,\eta)$ for which $\mathring{C}_x(\tau,\xi,\eta) = 0$ . If such a $(\tau,\xi,\eta)$ is an eigenvalue, i.e., $(\tau,\xi,\eta) \in \mathbb{C} \times \mathbb{C} \times \mathbb{R}$ , with $\operatorname{Im}\tau < 0$ , then $(\tau,\xi,\eta)$ corresponds to an evanescent wave. Since an evanescent wave decays exponentially to zero in the domain $[0,+\infty)$ , its influence is confronted in a neighborhood of the boundary $x = 0$ . If $(\tau,\xi,\eta)$ is a generalized eigenvalue, then $C_x(\tau,\xi,\eta) = 0$ , i.e., the wave *does not* travel into the space domain $[0,+\infty)$ . Thus the eigenvalues or generalized eigenvalues with $\mathring{C}_x = 0$ *do not* carry energy into the interior. Kreiss [Krei72] observed that even with the existence of the eigenvalues and generalized eigenvalues for which $\mathring{C}_x = 0$ , a strong estimate of the solution in terms of the prescribed data only is still possible if the solution in the interior region only is concerned, that is, if $[0,+\infty)$ is replaced by $[\delta,+\infty)$ in the estimate Eq.(2.1.2) for some $\delta > 0$ .

## 3.4 Reflection Coefficients

Suppose a monochromatic wave $e^{i(\tau t+\xi^- x+\eta y)}$ with $(\tau,\xi^-,\eta)\in\mathcal{V}$ is formed in the interior of $[0,+\infty)$ , and propagates at a group velocity $C_x^- < 0$ . When the wave front reaches the boundary $x=0$ , reflection occurs. The incident wave is balanced by a collection of monochromatic reflected waves, resulting in a solution of the form

$$e^{i(\tau t+\xi^- x+\eta y)} + \sum_{j=1}^{m^+} c_j^+ e^{i(\tau t+\xi_j^+ x+\eta y)} ,$$

where each signal $e^{i(\tau t+\xi^+ x+\eta y)}$ has $\mathring{C}_x(\tau,\xi_j^+,\eta) \geq 0$ , propagating into the domain $[0,+\infty)$ . In general, the incident wave is composed of all leftgoing monochromatic waves, and the general solution is given by

$$u = \sum_{j=1}^{m^-} c_j^- e^{i(\tau t+\xi_j^- x+\eta y)} + \sum_{j=1}^{m^+} c_j^+ e^{i(\tau t+\xi_j^+ x+\eta y)} . \tag{3.4.1}$$

The coefficients $c^\pm$ are connected by the boundary condition $B^k u = 0,\ k = 1,\dots,m^+$ . When the solution Eq.(3.4.1) is inserted into the boundary condition, the result is

$$N_B^+(\tau,\eta)c^+ = -N_B^-(\tau,\eta)c^- .$$

If the boundary operator $\{B^k\}$ is well-posed, the matrix $N_B^+(\tau,\eta)$ is invertible; it is therefore possible to solve for $c^+$ from

$$c^+ = R_B(\tau,\eta)c^-, \quad R_B(\tau,\eta) = -(N_B^+(\tau,\eta))^{-1}N_B^-(\tau,\eta), \tag{3.4.2}$$

where $R_B$ is said to be the *reflection matrix* associated with the boundary condition $\{B^k\}$. Thus well-posedness in the sense of Hersh is equivalent to the solvability of the reflected waves from the incident waves for all $(\tau, \eta) \in \mathcal{A}_0$. However, the coefficient of the reflected wave $c^+$ may become arbitrarily large as $\operatorname{Im} \tau \to 0$. Once again, we see that we require UKC, which guarantees the uniform solvability of $c^+$ from $c^-$ for all $(\tau, \eta) \in \bar{\mathcal{A}}_0$.

3.4.1 DEFINITION. *The reflection coefficient* $r_B$ *is defined by*

$$r_B = r_B(\tau, \eta) = \|R_B(\tau, \eta)\| .$$

3.4.2 COROLLARY. $\{B^k\}$ *is strongly well-posed if and only if* $r_B(\tau, \eta)$ *is defined and uniformly bounded for all* $(\tau, \eta) \in \bar{\mathcal{A}}_0$.

PROOF: Follows from Def.3.3.2. ☐

3.4.3 COROLLARY. $\{B^k\}$ *is a strongly well-posed perfectly absorbing boundary operator if and only if* $r_B(\tau, \eta) = 0$ *for all* $(\tau, \eta) \in \bar{\mathcal{A}}_0$.

PROOF: Follows from Thm.2.3.5. ☐

The physical interpretation of Cor.3.4.3 is obvious: if a boundary condition does not produce any reflection, it is a perfectly absorbing boundary condition. Also, because of Thm.2.3.6, a boundary condition with small reflection coefficient has a good absorption property.

# CHAPTER 4

# ABSORBING BOUNDARY CONDITIONS

In this chapter, we will develop a nonreflecting principle and construct from it a perfectly absorbing boundary operator. However, the perfectly absorbing boundary operator is not useful in practice because it is nonlocal in both time and space. It is therefore approximated by some local absorbing boundary operators. In doing so, one must satisfy a criterion that the reflection coefficient of the boundary operator should be small. The justification of this criterion is Thm.2.3.6. In section one, we develop a nonreflecting principle and a perfectly absorbing boundary operator. Section two deals with rational approximations of the perfectly absorbing boundary operator. In section three, we describe a process which characterizes the local absorbing boundary operators. Since the local boundary operators contain some optimization parameters, we discuss some strategies of selecting these parameters in section four.

In this chapter, we assume that the equation Eq.(2.1.1a) is homogeneous, and furthermore, that Assumption 2.2.1 holds.

## 4.1 Perfectly Absorbing Boundary Operators

We have shown in Thm.2.3.5 that for a boundary operator $\{B^k\}$ to be perfectly absorbing, it is necessary that

$$N_B^-(\tau,\eta) = \mathcal{O}\,, \quad \text{for all} \quad (\tau,\eta) \in \mathcal{A}_0\,.$$

Under the Assumption 2.2.1, the entries of $N_B^-$ are equal to $B^k(\tau,\xi_j^-,\eta)$. Our goal is therefore to find $\{B^k\}$ such that

$$B^k(\tau,\xi_j^-(\tau,\eta),\eta)=0\,,\quad k=1,\ldots,m^+,\quad j=1,\ldots,m^-\,,$$

for all $(\tau,\eta)\in\mathcal{A}_0$. As we have shown in chapter 3, each mode $(\tau,\xi_j^-,\eta)$ corresponds to a leftgoing monochromatic wave, as a result, $\mathring{C}_x(\tau,\xi_j^-,\eta)\le 0$ for all $(\tau,\eta)\in\mathcal{A}_0$. If we define

$$\mathring{Q}(\tau,\xi,\eta)=|\mathring{C}_x(\tau,\xi,\eta)|+\mathring{C}_x(\tau,\xi,\eta)\,, \tag{4.1.1}$$

then we have the desired property because

$$\mathring{Q}(\tau,\xi_j^-,\eta)=|\mathring{C}_x(\tau,\xi_j^-,\eta)|+\mathring{C}_x(\tau,\xi_j^-,\eta)=0\,,\quad \text{for all }\ (\tau,\eta)\in\mathcal{A}_0\,.$$

We may regard $\mathring{Q}$ of Eq.(4.1.1) as the symbol of a pseudo-differential operator as defined in Def.2.2.6, and apply it to the solution at the boundary $x=0$ to yield a boundary condition. This boundary condition, however, cannot be well-posed unless $m^+=1$, i.e., the equation Eq.(2.1.1a) admits only one rightgoing mode. For $m^+>1$, a well-posed boundary operator needs to be found, whose composition with $\mathring{Q}$ forms a well-posed perfectly absorbing boundary condition. Thus we assume $\{B^k\}$, $k=1,\ldots,m^+$, is some boundary operator, and define $\{B^k_{\mathring{Q}}\}$ by

$$B^k_{\mathring{Q}}(\tau,\xi,\eta)=\mathring{Q}(\tau,\xi,\eta)B^k(\tau,\xi,\eta)\,,\quad k=1,\ldots,m^+\,. \tag{4.1.2}$$

4.1.1 THEOREM. *For any well-posed $\{B^k\}$, the boundary operator $\{B^k_{\mathring{Q}}\}$ defined by Eq.(4.1.2) is well-posed and perfectly absorbing.*

PROOF: We will apply Thm.2.3.5. Let $\xi_j^+(\tau,\eta)$, $\xi_j^-(\tau,\eta)$ be the roots of the dispersion relation $P(\tau,\xi,\eta)=0$ for $(\tau,\eta)\in\mathcal{A}_0$, classified according to

Eq.(2.1.5). Then the $(k,j)$ -th entry of $N^-_{B_{\mathring{Q}}}(\tau,\eta)$ is equal to

$$[N^-_{B_{\mathring{Q}}}(\tau,\eta)]_{k,j} = \mathring{Q}(\tau,\xi^-_j(\tau,\eta),\eta)B^k(\tau,\xi^-_j(\tau,\eta),\eta) = 0\,, \quad \text{for all } (\tau,\eta)\in\mathcal{A}_0\,,$$

because $\mathring{C}_x(\tau,\xi^-_j(\tau,\eta),\eta) < 0$ , implying $\mathring{Q}(\tau,\xi^-_j(\tau,\eta),\eta) = 0$ . This shows that $N^-_{B_{\mathring{Q}}} = \mathcal{O}$ . Similarly, by examining each entry of $N^+_{B_{\mathring{Q}}}$ , we find that

$$\Delta^+_{B_{\mathring{Q}}}(\tau,\eta) = \Delta^+_B(\tau,\eta)\prod_{j=1}^{m^+}\mathring{Q}(\tau,\xi^+_j(\tau,\eta),\eta) \neq 0\,, \quad \text{for all } (\tau,\eta)\in\mathcal{A}_0\,,$$

because $\mathring{C}_x(\tau,\xi^+_j(\tau,\eta),\eta) > 0$ and $\Delta^+_B(\tau,\eta) \neq 0$ . Then this theorem follows from Thm.2.3.5. ☐

Thm.4.1.1 indicates that for any given (well-posed) boundary operator $\{B^k\}$ , the application of $\mathring{Q}$ is to totally eliminate the reflection caused by $\{B^k\}$ . For this reason, Thm.4.1.1 will be referred to as the ***nonreflecting principle*** and $\mathring{Q}$ will be called a *perfectly absorbing operator*.

The boundary operator $\{B^k_{\mathring{Q}}\}$ is not strongly well-posed even when $\{B^k\}$ is strongly well-posed, because the former always admits eigenvalues and generalized eigenvalues for which $\mathring{C}_x(\tau,\xi,\eta) = 0$ . However, such eigenvalues and generalized eigenvalues exert influence only in a neighborhood of the boundary, and they have little effect on the solution in the interior. See the discussion of § 3.3.

The operator $\mathring{Q}$ is only of theoretical importance, because its construction requires information about the roots of the dispersion relation. In the following, we will derive a weaker form of $\mathring{Q}$ which can be constructed explicitly without *a priori* information about the roots.

According to Def.2.2.6, the action of $\{B^k_{\mathring{Q}}\}$ on a function of the form

$$\sum c_j e^{(\tau t+\xi_j x+\eta y)}$$

is

$$B_Q^k u = \sum_j \iint \left\{ \left| \frac{\lambda}{\operatorname{Im}\xi_j} \right| - \frac{\lambda}{\operatorname{Im}\xi_j} \right\} B^k(\tau, \xi_j, \eta) e^{i(\tau t + \xi_j x + \eta y)} c_j(\tau, \eta) d\sigma d\eta \,,$$

for $k = 1, \ldots, m^+$ , where $\tau = \sigma + i\lambda$ , $\lambda < 0$ . The above equality holds independent of $\lambda$ , and it remains true in the limit $\lambda \to 0$ . That is

$$B_Q^k u = \lim_{\lambda \to -0} \sum_j \iint \left\{ \left| \frac{\lambda}{\operatorname{Im}\xi_j} \right| - \frac{\lambda}{\operatorname{Im}\xi_j} \right\} B^k(\tau, \xi_j, \eta) e^{i(\tau t + \xi_j x + \eta y)} c_j(\tau, \eta) d\sigma d\eta \,,$$

for $k = 1, \ldots, m^+$ .

If we formally interchange the order of limit and integration, we have, since $\lim_{\lambda \to -0} \mathring{C}_x = C_x$ ,

$$B_Q^k u = \sum_j \iint \left\{ \left| C_x(\tau, \xi_j, \eta) \right| + C_x(\tau, \xi_j, \eta) \right\} B^k(\tau, \xi_j, \eta) e^{i(\tau t + \xi_j x + \eta y)} c_j(\tau, \eta) d\tau d\eta \,,$$

for $k = 1, \ldots, m^+$ , where $\tau \in \mathbb{R}$ . This leads us to conjecture that an equivalent boundary operator of Eq.(4.1.2) is given by

$$B_Q^k u(\tau, \xi, \eta) = Q(\tau, \xi, \eta) B^k u \,, \quad k = 1, \ldots, m^+ \,, \tag{4.1.3}$$

where

$$Q(\tau, \xi, \eta) = |C_x(\tau, \xi, \eta)| + C_x(\tau, \xi, \eta) \,. \tag{4.1.4}$$

Unfortunately, this is not quite correct. Thm.3.2.1 says that $\lim_{\lambda \to 0} \mathring{C}_x(\tau, \xi, \eta) = C_x(\tau, \xi, \eta)$ holds only if $\operatorname{Im}\xi \to 0$ as $\lambda \to 0$ . That is, the limit of $(\tau, \xi, \eta)$ must be in $\mathcal{V}$ , corresponding to a travelling mode. Therefore, for the above argument to hold, it is necessary that $c_j(\tau, \eta) = 0$ for all $(\tau, \eta) \in \mathbb{R} \times \mathbb{R}$ such that $\operatorname{Im}\xi(\tau, \eta) \neq 0$ . This amounts to requiring that the solution consist of travelling waves only. When this condition is satisfied, we can prove the following theorem.

4.1.2 THEOREM. *Consider the IBVP Eq.(2.3.3). Assume* $\{A^k\}$ *and* $\{B^k\}$ *are strongly well-posed. Let* $\{B^k_Q\}$ *and* $Q$ *be defined as in Eq.(4.1.3) and Eq.(4.1.4) respectively. If* $g$ *generates travelling waves only, i.e., the Fourier transform* $\hat{g}$ *of* $g$ *satisfies*

$$supp\, \hat{g} \subset \Gamma = \{ \, (\tau,\eta) \in \mathbb{R} \times \mathbb{R} \bigm| P(\tau,\xi,\eta) \text{ has only real roots } \, \},$$

*then* $\{B^k_Q\}$ *is a perfectly absorbing boundary condition.*

PROOF: We need to show that $\{B^k_Q\}$ annihilates the solution of Eq.(2.3.2). By using the Fourier-Laplace transform, the solution of Eq.(2.3.2) is given by

$$u(t,x,y) = \iint_\Gamma [e^{i(\tau t+\xi_1^- x+\eta y)}, \dots , e^{i(\tau t+\xi_{m^-}^- x+\eta y)}] \cdot (N_A^-(\tau,\eta))^{-1} \hat{g}(\tau,\eta) d\tau d\eta \, .$$

When $B^k_Q$ is applied to the above expression according to Def.2.2.6, the result is

$$\begin{aligned} B^k_Q u = \iint_\Gamma [B^k(\tau,\xi_1^-,\eta) e^{i(\tau t+\xi_1^- x+\eta y)}, \dots ,B^k(\tau,\xi_{m^-}^-,\eta) e^{i(\tau t+\xi_{m^-}^- x+\eta y)}] \cdot \mathbb{Q}^- \\ \cdot (N_A^-)^{-1} \hat{g}(\tau,\eta) d\tau d\eta \, , \end{aligned}$$

where $\mathbb{Q}^-$ is an $m^- \times m^-$ diagonal matrix given by

$$\mathbb{Q}^- = \operatorname{diag} \{ Q(\tau,\xi_1^-,\eta), \dots , Q(\tau,\xi_{m^-}^-,\eta) \} \, .$$

For each $j = 1, \dots , m^-$, we have $C_x(\tau,\xi_j^-,\eta) \leq 0$ according to Thm.3.2.1. Therefore

$$Q(\tau,\xi_j^-,\eta) = |C_x(\tau,\xi_j^-,\eta)| + C_x(\tau,\xi_j^-,\eta) = 0 \, ,$$

for all $(\tau,\eta) \in \Gamma$. This shows that $\{B^k_Q\}$ annihilates the solution of Eq.(2.3.2) and the proof is complete. □

As the above theorem shows, the boundary operator $Q$ defined in Eq.(4.1.4) annihilates only travelling wave solutions. If the interior disturbance consists of both travelling and evanescent waves, $Q$ is no longer perfectly absorbing. However this is of little consequence, because evanescent waves decay exponentially in space, so that their reflection can be effectively reduced if the artificial boundary is placed some distance away from the initial disturbance.

An important advantage of $Q$ over the perfectly absorbing operator $\mathring{Q}$ is that $Q$ can be explicitly constructed from the dispersion relation $P(\tau,\xi,\eta) = 0$ without knowledge of its roots. From Eq.(3.1.3), the $x$-component $C_x$ of group velocity is given by

$$C_x(\tau,\xi,\eta) = -\frac{\partial\tau}{\partial\xi}.$$

By the implicit function theorem, the above expression is equivalent to

$$C_x(\tau,\xi,\eta) = \frac{P_\xi}{P_\tau}.$$

Thus an explicit formula for $Q$ is given by

$$Q(\tau,\xi,\eta) = \left|\frac{P_\xi(\tau,\xi,\eta)}{P_\tau(\tau,\xi,\eta)}\right| + \frac{P_\xi(\tau,\xi,\eta)}{P_\tau(\tau,\xi,\eta)}. \tag{4.1.5}$$

In order for Eq.(4.1.5) to be defined for all $(\tau,\xi,\eta) \in \mathcal{V}$, we make the following assumption.

4.1.3 ASSUMPTION. *We assume that the hyperbolic equation Eq.(1.1.1a) is strictly hyperbolic [Duff76, p50], i.e., for any* $(\xi,\eta) \in \mathbb{R} \times \mathbb{R}^{n-1}$, $\xi^2 + |\eta|^2 \neq 0$, *the equation*

$$P(\tau,\xi,\eta) = 0$$

*has only real and distinct roots* $\tau$.

Therefore, if the equation Eq.(2.1.1a) is strictly hyperbolic, then $P_\tau(\tau,\xi,\eta) \neq 0$ for all $(\tau,\xi,\eta) \in \mathcal{V} \setminus (0,0,0)$. Thus $Q(\tau,\xi,\eta)$ in Eq.(4.1.5) is well defined for all $(\tau, \xi, \eta) \in \mathcal{V} \setminus (0,0,0)$ .

## 4.2 Rational Approximations

In the previous section, we derived a perfectly absorbing operator given by Eq.(4.1.5). Because of the presence of the absolute value function in its definition, the operator $Q$ of Eq.(4.1.5) is not a differential operator; rather it is a pseudo-differential operator defined in Def.2.2.6. Since a pseudo-differential operator is nonlocal in both space and time, $Q$ is not useful as far as numerical implementation is concerned.

In order to obtain a local boundary operator approximating Eq.(4.1.5), the absolute value function must be approximated by a rational function. By making a rational transform if necessary, we may assume that $C_x \in [-1,1]$ .

Let $r(x) = \dfrac{p_m(x)}{q_n(x)}$ of order $(m,n)$ be a rational approximation to $|x|$ on the interval $[-1,1]$ , then Eq.(4.1.5) can be approximated by

$$r(\frac{P_\xi}{P_\tau}) + \frac{P_\xi}{P_\tau} = \frac{p_m(P_\xi/P_\tau)}{q_n(P_\xi/P_\tau)} + \frac{P_\xi}{P_\tau} .$$

Thus $Q(\tau, \xi, \eta) = 0$ can be approximated by

$$\frac{p_m(P_\xi/P_\tau)}{q_n(P_\xi/P_\tau)} + \frac{P_\xi}{P_\tau} = 0 .$$

Multiplying both sides of the above equation by $q_n$, we have

$$p_m(\frac{P_\xi}{P_\tau}) + q_n(\frac{P_\xi}{P_\tau})\,\frac{P_\xi}{P_\tau} = 0 , \qquad (4.2.1)$$

and multiplying both sides by $P_\tau^d$, we obtain

$$P_\xi P_\tau^{d-1} q_n(\frac{P_\xi}{P_\tau}) + P_\tau^d p_m(\frac{P_\xi}{P_\tau}) = 0 , \qquad (4.2.2)$$

where $d = \max\{m, n+1\}$ . The left hand side of Eq.(4.2.2) is a homogeneous polynomial of degree $d \times (p-1)$ , where $p$ is the degree of $P$ . Hence it is the symbol of a differential operator.

An obvious choice of $r(x)$ seems to be the function, $r^*(x)$ , obtained by Newman [Newm64], which approximates $|x|$ on $[-1,1]$ with an accuracy of order $e^{-c\sqrt{d}}$ . This approximation, however, results in a boundary condition that is not strongly well-posed, because $r^*(0) = 0$ ; hence, Eq.(4.2.2) admits a generalized eigenvalue. Although adding a small positive constant to $r^*$ would rule out all generalized eigenvalues while still maintaining roughly the same accuracy, it is not clear whether the resulting boundary condition admits other eigenvalues. Rather than looking for a function approximating $|x|$ , we will instead take another approach in which we study the necessary form of a strongly well-posed boundary operator with the smallest reflection coefficient.

The left hand side of Eq.(4.2.1) can be expressed as

$$Q_d(\frac{P_\xi}{P_\tau}),$$

where $Q_d$ is a polynomial of degree $d = \max\{m, n+1\}$ with real coefficients, which can be factorized as

$$Q_d(\frac{P_\xi}{P_\tau}) = \prod_{j=1}^{d_1} (\frac{P_\xi}{P_\tau} - v_j) \prod_{j=d_1+1}^{\frac{d+d_1}{2}} [(\frac{P_\xi}{P_\tau} - w_j)^2 + \hat{w}_j^2], \qquad (4.2.3)$$

with $v_j, w_j, \hat{w}_j \in \mathbb{R}$ . For any polynomial Eq.(4.2.3), there exists a corresponding differential operator whose symbol is given by

$$P_\tau^d Q_d(\frac{P_\xi}{P_\tau}). \qquad (4.2.4)$$

In order to avoid using too many notations, the boundary operator which has the symbol given by Eq.(4.2.4) will also be denoted by $Q_d$ . Thus when we say boundary operator $Q_d$, we in fact mean the differential operator defined by Eq.(4.2.4) (i.e., $P_\tau^d Q_d$) rather than the pseudo-differential operator defined by Eq.(4.2.3)[2].

To form an absorbing boundary condition at the boundary $x = 0$, we assume that there exists a strongly well-posed boundary operator $\{B^k\}$, e.g., $B^k(\tau, \xi, \eta) = \xi^{k-1}$ under Assumption 2.2.1, whose composition with $Q_d$ yields

$$B_{Q_d}^k u = Q_d B^k u = 0\,. \tag{4.2.5}$$

The reflection matrix of the boundary condition Eq.(4.2.5) is given by

$$R_{Q_d} = -(N_{B_{Q_d}}^+)^{-1} N_{B_{Q_d}}^-$$

where

$$N_{B_{Q_d}}^+ = N_B^+ \mathbb{Q}_d^+\,, \quad N_{B_{Q_d}}^- = N_B^- \mathbb{Q}_d^-\,.$$

In the above, $\mathbb{Q}_d^\pm$ is an $m^\pm \times m^\pm$ diagonal matrix whose elements are given by

$$[\mathbb{Q}_d^\pm]_{j,j} = P_\tau^d(\tau, \xi_j^\pm, \eta) Q_d(\tau, \xi_j^\pm, \eta)\,, \quad j = 1, \ldots, m^\pm\,.$$

According to Def.3.4.1, the reflection coefficient can be estimated by

$$r_{Q_d} = \|R_{Q_d}\| \le \|(\mathbb{Q}_d^+)^{-1}\| \|\mathbb{Q}_d^-\| \|(N_B^+)^{-1}\| \|N_B^-\|\,. \tag{4.2.6}$$

4.2.1 DEFINITION. *The absorbing factor of the boundary operator Eq.(4.2.4) is defined by*

$$a_{Q_d}(\tau, \eta) = \|(\mathbb{Q}_d^+)^{-1}(\tau, \eta)\| \|\mathbb{Q}_d^-(\tau, \eta)\|\,.$$

---

[2] However, the boundary conditions $Q_d u = 0$ and $P_\tau^d Q_d u = 0$ are not equivalent so far as the reflection matrices are concerned, unless $P_\tau(\tau, \xi, \eta)$ is independent of $\xi$. This abuse of notation should not cause any confusion because in this section, we are only interested in the *differential operators*.

In the following, we will only be concerned with the operator $Q_d$ because $a_{Q_d}$ is the only factor in Eq.(4.2.6) that is relevant to the rational approximation. Our goal is to find a polynomial $Q_d$ which yields a well-posed operator Eq.(4.2.4) with the smallest absorbing factor $a_{Q_d}$ .

4.2.2 LEMMA. *Let the polynomial* $Q_d$ *be given by Eq.(4.2.3). Then the corresponding absorbing factor* $a_{Q_d}$ *of Eq.(4.2.4) is given by*

$$a_{Q_d} = \prod_{j=1}^{d_1} \frac{P_\xi(\xi_0^-) - v_j P_\tau(\xi_0^-)}{P_\xi(\xi_0^+) - v_j P_\tau(\xi_0^+)} \times \prod_{j=d_1+1}^{\frac{d+d_1}{2}} \frac{\left\{P_\xi(\xi_0^-) - w_j P_\tau(\xi_0^-)\right\}^2 + \hat{w}_j^2 P_\tau^2(\xi_0^-)}{\left\{P_\xi(\xi_0^+) - w_j P_\tau(\xi_0^+)\right\}^2 + \hat{w}_j^2 P_\tau^2(\xi_0^+)}, \qquad (4.2.7)$$

*where* $\xi_0^\pm = \xi_j^\pm$ *for some* $j = 1, \ldots, m^\pm$ .

PROOF: The result follows from Def.4.2.1 by noting that

$$\begin{aligned} \|(Q_d^+)^{-1}\| &= \left\{\min_j \{P_\tau^d Q_d(\xi_j^+)\}\right\}^{-1} = \left\{P_\tau^d Q_d(\xi_0^+)\right\}^{-1} \\ \|Q_d^-\| &= \max_j \{P_\tau^d Q_d(\xi_j^-)\} \qquad = P_\tau^d Q_d(\xi_0^-) . \end{aligned}$$

□

Now, we assume $C_x = \dfrac{P_\xi}{P_\tau}(\tau, \eta) \in [\alpha, \beta]$ for all $(\tau, \eta) \in \Gamma$ and $-\infty < \alpha < 0 < \beta < +\infty$ ; that is, the $x$ -component of group velocity is bounded above and below by $\alpha$ and $\beta$ , respectively.

4.2.3 DEFINITION. *Given* $\varepsilon : 0 < \varepsilon < -\alpha$ , *let*

$$\Pi_d = \{\ Q_d \ |Q_d \text{ is a real polynomial of degree } d\ \} .$$

$$\Gamma_\varepsilon = \{\ (\tau, \eta) \in \Gamma \ |\ \alpha \le C_x(\tau, \xi^-(\tau, \eta), \eta) \le -\varepsilon\ \} .$$

$$\Gamma_0 = \{\ (\tau, \eta) \in \Gamma \ |\ \alpha \le C_x(\tau, \xi^-(\tau, \eta), \eta) \le 0\ \} .$$

*A boundary operator* $Q_d^*$ *is said to be an* $\mathcal{L}_\varepsilon^\infty$ *-best absorbing boundary operator of degree* $d$ *if*

$$\max_{(\tau,\eta)\in\Gamma_\varepsilon} a_{Q_d^*}(\tau,\eta) = \min_{Q_d\in\Pi_d}\Big\{\max_{(\tau,\eta)\in\Gamma_\varepsilon} a_{Q_d}(\tau,\eta)\Big\}.$$

$\mathcal{L}_\varepsilon^\infty$ -best absorbing boundary operators are the optimal boundary operators of a fixed degree, in the sense that they minimize the reflection coefficient over all travelling waves of propagating speed in the range $[\alpha,-\varepsilon]$. In applications, during the time interval of interest, the waves with slow speed $(C_x \in (-\varepsilon, 0])$ will not travel far enough to reach the boundary.

The following definition, due to Wagatha [Waga83], takes into consideration wave modes of every speed in $[\alpha, 0)$.

4.2.4 DEFINITION. *Let* $\rho \in \mathcal{L}^1_{\Gamma_\alpha}$ *with* $\rho \geq 0$ *and*

$$\iint_{\Gamma_0} \rho(\tau,\eta)d\tau d\eta = 1.$$

*A boundary operator* $Q_d^*$ *is said to be an* $\mathcal{L}^1$ *-best absorbing boundary operator of degree* $d$ *if*

$$\iint_{\Gamma_0} \rho(\tau,\eta)a_{Q_d^*}(\tau,\eta)d\tau d\eta = \min_{Q_d\in\Pi_d} \iint_{\Gamma_0} \rho(\tau,\eta)a_{Q_d}(\tau,\eta)d\tau d\eta.$$

*The* $\mathcal{L}^2$ *-best absorbing boundary operator is similarly defined by replacing* $a_{Q_d}$ *by* $a^2_{Q_d}$ *in the above integral.*

An immediate consequence of the above definitions is

4.2.5 THEOREM. *The* $\mathcal{L}_\varepsilon^\infty$ *-,* $\mathcal{L}^1$ *- and* $\mathcal{L}^2$ *-best absorbing boundary operators exist.*

PROOF: This follows from the fact that $\Pi_d$ is a finite dimensional space. □

## 4.3 Characterization of Best Absorbing Operators

In this section, we will describe a factorization process which characterizes general absorbing boundary operators. We will show that the best absorbing boundary operators defined in the previous section must be in a canonical form; they can be factorized into a product of some $(p-1)$ st order differential ~~operators,~~ where $p$ is the order of the differential operator $P$.

4.3.1 THEOREM. *Let $\varepsilon > 0$. If $Q_d^* \in \Pi_d$ defines a strongly well-posed $\mathcal{L}_\varepsilon^\infty$-best absorbing boundary operator of degree $d$, then*

$$Q_d^*(x) = \prod_{j=1}^{d} (x - v_j), \tag{4.3.1}$$

*with $v_j \in [\alpha, -\varepsilon]$.*

PROOF: Since $Q_d^*$ is a polynomial of degree $d$, it can be factorized as

$$Q_d^*(x) = \prod_{j=1}^{d_1} (x - v_j) \prod_{j=d_1+1}^{\frac{d+d_1}{2}} \left[(x - w_j)^2 + \hat{w}_j^2\right]. \tag{4.3.2}$$

Let $a^* = a_{Q_d^*}$ be the absorbing factor of the boundary operator $Q_d^*$ as defined in Eq.(4.2.4). According to Lem.4.2.2, $a^*$ is given by Eq.(4.2.7).

We first show that if $Q_d^*$ is an $\mathcal{L}_\varepsilon^\infty$-best absorbing boundary operator, then $d_1 = d$ in Eq.(4.3.2), i.e., the polynomial $Q_d^*$ contains no quadratic factors. Thus we consider a quadratic factor $(x - w_j)^2 + \hat{w}_j^2$ and its corresponding factor in Eq.(4.2.7)

$$\begin{aligned} a_j^* &= \left| \frac{(P_\xi(\xi_0^-) - w_j P_\tau(\xi_0^-))^2 + \hat{w}_j^2 P_\tau^2(\xi_0^-)}{(P_\xi(\xi_0^+) - w_j P_\tau(\xi_0^+))^2 + \hat{w}_j^2 P_\tau^2(\xi_0^+)} \right| \\ &= \hat{a}_j^* c^2 , \end{aligned}$$

where

$$\hat{a}_j^* = \frac{(x^- - w_j)^2 + \hat{w}_j^2}{(x^+ - w_j)^2 + \hat{w}_j^2},$$

and $c = \dfrac{P_\tau(\xi_0^-)}{P_\tau(\xi_0^+)}$ , $x^\pm = \dfrac{P_\xi}{P_\tau}(\xi_0^\pm) = C_x(\tau, \xi_0^\pm, \eta)$ . By assumption,

$$\alpha \le x^- \le -\varepsilon < 0 \le x^+ \le \beta .$$

It is easy to verify, by expanding both sides, that

$$\left(x^+ + \sqrt{w_j^2 + \hat{w}_j^2}\right)^2 \ge (x^+ - w_j)^2 + \hat{w}_j^2$$

and

$$\left(x^- + \sqrt{w_j^2 + \hat{w}_j^2}\right)^2 \le (x^- - w_j)^2 + \hat{w}_j^2 .$$

This implies

$$\hat{a}_j^* \ge \frac{\left(x^- + \sqrt{w_j^2 + \hat{w}_j^2}\right)^2}{\left(x^+ + \sqrt{w_j^2 + \hat{w}_j^2}\right)^2} \quad \text{for all} \quad x^+,\ x^- .$$

The equality in the above holds only if $\hat{w}_j = 0$ . Because $Q_d^*$ is optimal, the value of $\hat{a}_j^*$ cannot be reduced, hence one concludes $\hat{w}_j = 0$ . In that case, this quadratic factor is in fact a product of two linear factors. This shows that Eq.(4.3.1) holds.

We now consider a linear factor $(x - v_j)$ of Eq.(4.3.1) and show that if $v_j \notin [\alpha, -\varepsilon]$ , then either the boundary operator Eq.(4.3.1) fails the strong well-posedness criterion, or it is not optimal.

If $v_j \in [0, \beta]$ , then the boundary operator Eq.(4.3.1) is not strongly well-posed because there exists a generalized eigenvalue with $C_x(\tau, \xi^+, \eta) = v_j$ . Now we consider the reflection coefficient corresponding to this factor,

$$a_j^* = \left|\frac{x^- - v_j}{x^+ - v_j}\right| c ,$$

where $x^{\pm}$ and $c$ are as defined before.

It is easy to verify that if $v_j \in (-\infty, \alpha) \cup (\beta, +\infty)$ , then

$$\left| \frac{x^- - v_j}{x^+ - v_j} \right| > \left| \frac{x^- - \alpha}{x^+ - \alpha} \right|, \qquad \text{for all} \quad x^+,\ x^- .$$

Also, if $v_j \in (-\varepsilon, 0)$ , then

$$\left| \frac{x^- - v_j}{x^+ - v_j} \right| > \left| \frac{x^- + \varepsilon}{x^+ + \varepsilon} \right|, \qquad \text{for all} \quad x^+,\ x^- .$$

Therefore $v_j \in [\alpha, -\varepsilon]$ for $Q_d^*$ to be optimal. This completes the proof. ☐

4.3.2 THEOREM. *If $Q_d^* \in \Pi_d$ defines a strongly well-posed $\mathcal{L}^1$ - (or $\mathcal{L}^2$ -) best absorbing boundary operator of degree $d$ , then Eq.(4.3.1) holds and $v_j \in [\alpha, 0)$ .*

PROOF: This is identical to Thm.4.3.1. ☐

From the above theorems, the strongly well-posed $\mathcal{L}_\varepsilon^\infty$ - ( $\mathcal{L}^1$ - or $\mathcal{L}^2$ -) best absorbing operator of degree $d$ must have the form

$$Q_d^*(D_t, D_x, D_y) = \prod_{j=1}^{d} \Big\{ P_\xi(D_t, D_x, D_y) - v_j P_\tau(D_t, D_x, D_y) \Big\}, \tag{4.3.3}$$

where $v_j \in [\alpha, -\varepsilon]$ . Each factor is a differential operator of order $p-1$ , where $p$ is the order of the equation Eq.(2.1.1a). $v_j$ remains to be determined.

4.3.3 THEOREM. *Let $Q_d^*$ be an $\mathcal{L}_\varepsilon^\infty$ -best absorbing boundary operator and*

$$Q_d^*(x) = \prod_{j=1}^{d} (x - v_j) .$$

*Then*

$$v_j \neq v_k , \qquad \text{if } j \neq k .$$

**PROOF:** We show that if there is a factor $(x - v_j)^2$ in $Q_d^*$, then $Q_d^*$ cannot be optimal. Thus consider a factor of $a^* = a_{Q_d^*}$ given by Eq.(4.2.7)

$$a_j^* = \frac{(x^- - v_j)^2}{(x^+ - v_j)^2} c^2 ,$$

where $c$, which is independent of $v_j$, is defined in the proof of Thm.4.3.1. We consider a polynomial $\hat{Q}$ obtained by replacing the factor $(x - v_j)^2$ in $Q_d^*$ by

$$(x - v_j - \delta)(x - v_j + \frac{\delta v_j}{v_j + \delta}) ,$$

for some $\delta$, $0 < \delta < -v_j$. The relevant factor in $a_{\hat{Q}}$ is given by

$$\begin{aligned} \hat{a}_j &= \frac{\left|(x^- - v_j - \delta)(x^- - v_j + \frac{\delta v_j}{v_j+\delta})\right|}{\left|(x^+ - v_j - \delta)(x^+ - v_j + \frac{\delta v_j}{v_j+\delta})\right|} c^2 \\ &= \frac{\left|(x^-)^2 - (v_j + \delta + \frac{v_j^2}{v_j+\delta})x^- + v_j^2\right|}{\left|(x^+)^2 - (v_j + \delta + \frac{v_j^2}{v_j+\delta})x^+ + v_j^2\right|} c^2 . \end{aligned}$$

The denominator of $\hat{a}_j$ is an increasing function of $\delta$ for all $x^+ \geq 0$ while the numerator is decreasing if

$$x^- \notin [\frac{v_j^2}{v_j + \delta}, v_j + \delta] .$$

Thus $\hat{a}_j = a_j^*$ if $\delta = 0$, and

$$\hat{a}_j < a_j^* \qquad \text{if } x^- \notin [\frac{v_j^2}{v_j + \delta}, v_j + \delta] \tag{4.3.4}$$

for all $\delta$, with $0 < \delta < -v_j$. Define

$$M = \max_{\substack{0 \leq x^+ \leq \beta \\ \alpha \leq x^- \leq -\epsilon}} a_j^*(x^+, x^-) .$$

Since $\hat{a}_j = 0$, if both $\delta = 0$ and $x^- = v_j$, there exists a $\delta > 0$ such that

$$\hat{a}_j(x^+, x^-) \leq \frac{M}{2}, \qquad \text{if } x^- \in [\frac{v_j^2}{v_j + \delta}, v_j + \delta] . \tag{4.3.5}$$

The combination of Eq.(4.3.4) and Eq.(4.3.5) yields

$$\hat{a}_j < a_j^* \le M\,, \quad \text{for all } x^+ \in [0,\beta], \quad x^- \in [\alpha,-\varepsilon] \setminus [\frac{v_j^2}{v_j+\delta}, v_j+\delta]$$

$$\hat{a}_j \le \frac{M}{2}\,, \quad \text{for all } x^+ \in [0,\beta], \quad x^- \in [\frac{v_j^2}{v_j+\delta}, v_j+\delta]\,.$$

Consequently $\hat{a}_j < M$ for all $(x^+, x^-) \in [0,\beta] \times [\alpha,-\varepsilon]$ . This shows that $Q_d^*$ is not optimal. ☐

A question to be asked is whether the absorption property of the absorbing boundary operator Eq.(4.3.3) would improve if the degree $d$ of $Q_d^*$ is increased. According to Thm.2.3.6, if $a_{Q_d^*}$ tends to zero as $d \to \infty$ , then the operator Eq.(4.3.3) approaches a perfectly absorbing boundary operator (in the sense of Thm.2.3.6). Therefore, the following theorem answers the question we just raised.

4.3.4 THEOREM. *Suppose*

$$\max_j \left|P_\tau(\tau,\xi_j^-,\eta)\right| \le \min_j \left|P_\tau(\tau,\xi_j^+,\eta)\right|$$

*for all* $(\tau,\eta) \in \Gamma_\varepsilon$ . *If* $Q_d^*$ *is an* $\mathcal{L}_\varepsilon^\infty$ *-best absorbing boundary operator, then* $a_{Q_d^*}(\tau,\eta) < 1$ *for all* $(\tau,\eta) \in \Gamma_\varepsilon$ *and*

$$\lim_{d\to\infty} a_{Q_d^*} = 0 \quad \textit{for all } (\tau,\eta) \in \Gamma_\varepsilon\,.$$

PROOF: Let

$$\hat{Q}_d(x) = \prod_{j=1}^{d} (x - 2\alpha),$$

then $a_{\hat{Q}_d} < 1$ and $\lim_{d\to+\infty} a_{\hat{Q}_d} = 0$ . The theorem follows from observing that $a_{Q_d^*} \le a_{\hat{Q}_d}$ . ☐

We have shown that a strongly well-posed best absorbing boundary operator is necessarily of the form Eq.(3.3.3). However, it is not clear whether the operator Eq.(3.3.3) is strongly well-posed (or well-posed for that matter). The well-posedness of Eq.(3.3.3) needs to be determined for each individual case, but the following theorem renders a mechanism to simplify this process.

Eq.(4.3.3) can be applied to any strongly well-posed boundary operator $\{B^k\}$ to yield an absorbing boundary operator. Thus we assume a strongly well-posed condition $\{B^k\}$ is given. Its composition with Eq.(4.3.3) has a symbol

$$B^k_{Q_d^*}(\tau,\xi,\eta) = \prod_{j=1}^{d} (P_\xi(\tau,\xi,\eta) - v_j P_\tau(\tau,\xi,\eta)) B^k(\tau,\xi,\eta)\,, \; k = 1,\dots,m^+ \,. \tag{4.3.6}$$

4.3.5 THEOREM. *Suppose that Assumption 2.2.1 hold. Let* $\{B^k\}$ *be strongly well-posed and* $\{B^k_{Q_d^*}\}$ *be defined by Eq.(4.3.6). Then* $\{B^k_{Q_d^*}\}$ *is well-posed (strongly well-posed, resp.) if and only if for each* $j = 1,\dots,d$ ,

$$P_\xi(\tau,\xi,\eta) - v_j P_\tau(\tau,\xi,\eta) = 0 \tag{4.3.7a}$$

*and*

$$P(\tau,\xi,\eta) = 0 \tag{4.3.7b}$$

*have no common solutions* $(\tau,\xi,\eta) \in \mathbb{C} \times \mathbb{C} \times \mathbb{R}^{n-1}$ *for which* $(\tau,\eta) \in \mathcal{A}_0$ *and* $\mathring{C}_x(\tau,\xi,\eta) > 0$ *(* $(\tau,\eta) \in \bar{\mathcal{A}}_0$ *and* $\mathring{C}_x(\tau,\xi,\eta) \geq 0$ *, resp.).*

PROOF: According to Cor.2.2.3 (Def.3.3.2, resp.), the boundary condition $\{B^k_{Q_d^*}\}$ is well-posed (strongly well-posed, resp.) if and only if

$$\Delta^+_{Q_d^*} \neq 0\,, \quad \text{for all } (\tau,\eta) \in \mathcal{A}_0 \quad ((\tau,\eta) \in \bar{\mathcal{A}}_0\,,\ \text{resp.})\,. \tag{4.3.8}$$

A direct calculation shows that

$$\Delta^+_{Q_d^*} = \Delta^+_B \prod_{k=1}^{m^+} Q_d^*(\tau,\xi_k^+,\eta)\,,$$

where

$$Q^*_d(\tau,\xi_k^+,\eta) = \prod_{j=1}^{d} \left( P_\xi(\tau,\xi_k^+,\eta) - v_j P_\tau(\tau,\xi_k^+,\eta) \right) .$$

For each $k = 1,\dots,m^+$ , $(\tau,\xi_k^+,\eta)$ is a solution of $P(\tau,\xi,\eta) = 0$ . Furthermore, $\mathring{C}_x(\tau,\xi_k^+,\eta) > 0$ if $(\tau,\eta) \in \mathcal{A}_0$ and $\mathring{C}_x(\tau,\xi_k^+,\eta) \geq 0$ if $(\tau,\eta) \in \bar{\mathcal{A}}_0$ . The theorem follows from Eq.(4.3.8). ☐

Note that the boundary operator Eq.(4.3.6) is not strongly well-posed even when $\{B^k\}$ is strongly well-posed, because Eq.(4.3.6) admits a generalized eigenvalue $(\tau,\xi,\eta) = (0,0,0)$ . This is unavoidable for any homogeneous boundary conditions. The effect of this generalized eigenvalue will be discussed in Chapter 6. However, the boundary operator Eq.(4.3.6) does not admit any other generalized eigenvalues if $v_j \in [\alpha, -\varepsilon]$ , as guaranteed by Thm.4.3.1. Because of this, in later chapters, when we determine the well-posedness of a boundary condition, we need only to consider eigenvalues.

## 4.4 Selection of Parameters

The performance of the absorbing boundary operator Eq.(4.3.3) depends upon the choice of the parameters $v_j$ . The parameters of the $\mathcal{L}^\infty_\varepsilon$ -best absorbing boundary operators can be determined by minimizing the expression

$$\max_{\substack{x^- \in [\alpha,-\varepsilon] \\ x^+ \in [0,\beta]}} \prod_{j=1}^{d} \left| w_j(x^+,x^-) \frac{x^- - v_j}{x^+ - v_j} \right| . \qquad (4.4.1)$$

The weight $w_j$ is dependent on the dispersion relation $P(\tau,\xi,\eta) = 0$ and its roots. This shows that the best parameters are problem dependent; their determination requires knowledge of the solution of the differential equation. Nonetheless, we will propose some strategies where the parameters can be selected to produce highly absorbing boundary operators.

There is a natural physical interpretation of the parameters $v_j$ in Eq.(4.3.3); they are related to the group velocities of wave solutions. This connection is apparent from Eq.(4.4.1). The $j$ th factor in Eq.(4.4.1) vanishes at $x^- = v_j$. Since $x^- = \dfrac{P_\xi(\tau,\xi^-,\eta)}{P_\tau(\tau,\xi^-,\eta)} = C_x(\tau,\xi^-,\eta)$ is the group velocity associated with the monochromatic wave $e^{i(\tau t+\xi^- x+\eta y)}$, it can be concluded that the $j$ th factor of $Q_d^*$ annihilates the plane wave solution travelling at the group velocity $v_j$. Thus we have

4.4.1 PROPOSITION. *Let* $Q_j$ *be the* $j$ *th factor*

$$Q_j = P_\xi(D_t, D_x, D_y) - v_j P_\tau(D_t, D_x, D_y)$$

*of an absorbing boundary operator of degree* $d$. *If* $u = e^{i(\tau t+\xi x+\eta y)}$ *satisfies* $C_x(\tau,\xi,\eta) = v_j$, *then* $Q_j u = 0$.

PROOF:

$$Q_j u = \{P_\xi(\tau,\xi,\eta) - v_j P_\tau(\tau,\xi,\eta)\}u = P_\tau(\tau,\xi,\eta)(C_x(\tau,\xi,\eta) - v_j)u = 0.$$

☐

In general, if

$$u = \sum_{j=1}^{d} e^{i(\tau t+\xi_j x+\eta y)},$$

with $C_x(\tau,\xi_j,\eta) = v_j$, then

$$Q_d u = \prod_{j=1}^{d} Q_j u = 0.$$

This leads to our first strategy of determining the parameters. If, as *a priori* information, the solution is known to be a wave train concentrated at a finite number of modes $(\tau_j,\xi_j,\eta_j)$, then the parameters may be tuned to the propagation speeds of these modes, i.e., $v_j = C_x(\tau_j, \xi_j, \eta_j)$ , so that the resulting boundary operator produces only reflections of very low amplitude.

In the case where no such information is available or the solution consists of a fairly large spectrum of frequencies and wave numbers, one appeals again to Eq.(4.4.1), with some simplifications.

Suppose the weight $w_j$ is bounded and $v_j \in [\alpha, -\varepsilon]$ . Then the value in Eq.(4.4.1) can be estimated by

$$\max \prod_{j=1}^{d} \left| w_j \frac{x^- - v_j}{x^+ - v_j} \right| \le M \max_{x^- \in [\alpha, -\epsilon]} \prod_{j=1}^{d} \left| x^- - v_j \right| ,$$

where $M > 0$ is independent of $v_j$ . This leads to the our second strategy. The parameters $v_j$ 's are chosen in order to minimize the right hand side of the above inequality. The solution to this problem is explicitly knwon; $v_j$ 's are the zeros of the shifted Chebyshev polynomial of degree $d$ (see Chapter 7).

As Wagatha [Waga83] pointed out, $\mathcal{L}^1$ - or $\mathcal{L}^2$ -best boundary operators have certain advantages. The construction of such boundary operators requires $v_j$ be chosen to minimize the integral of Def.4.2.4. By a simplification procedure similar to the one adopted earlier, $v_j$ can be selected by minimizing instead the integral

$$\int_{\alpha}^{-\epsilon} \int_{0}^{\beta} \rho(x^+, x^-) \prod_{j=1}^{d} \left| \frac{x^- - v_j}{x^+ - v_j} \right| dx^+ dx^- ,$$

for some given weight function $\rho$ .

Examples of selecting parameters by using the above strategies will be given in the next chapter, where we study absorbing boundary conditions for some equations arising in applications.

# CHAPTER 5

# SOME SECOND-ORDER HYPERBOLIC EQUATIONS

In this chapter, we consider a few examples of second-order equations of physical interest, and derive absorbing boundary conditions for them. One equation which will be studied in detail is the linearized small disturbance equation of unsteady transonic flows. This equation is of substantial interest, and it is sufficiently complex to make good use of our general theory. Our theory applied to this example yields some simplifications and generalizations of boundary conditions established earlier for this example by Engquist and Majda [EnMa80].

While the theory developed in the last chapter can be applied directly to second-order equations without modifications, some results can in fact be simplified. Since $P(\tau, \xi, \eta)$ is now a polynomial of degree two, we may assume that there is exactly one root $\xi^+$ with $\operatorname{Im} \xi^+ \geq 0$ and another root $\xi^-$ with $\operatorname{Im} \xi^- \leq 0$. With this assumption, the absorbing boundary operator $Q_d$ obtained in Chapter 4 can be directly used for the boundary condition. That is, the $\mathcal{L}_\epsilon^\infty$ -best absorbing boundary condition has the form

$$Q_d u = \prod_{j=1}^{d} \{ P_\xi(D_t, D_x, D_y) - v_j P_\tau(D_t, D_x, D_y) \} u = 0 .$$

## 5.1 The Wave Equation and the Klein-Gordon Equation

Recall that the multidimensional wave equation is given by

$$u_{tt} = \triangle u ,$$

where $\triangle$ is the Laplace operator. Its dispersion relation is

$$P(\tau,\xi,\eta) = \tau^2 - (\xi^2 + |\eta|^2) = 0 .$$

A simple calculation shows that the $\mathcal{L}_\varepsilon^\infty$ -best absorbing boundary condition is of the form

$$Q_d\Big(\frac{\partial}{\partial t}, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}\Big)u = \prod_{j=1}^{d} \Big(\frac{\partial}{\partial x} + v_j \frac{\partial}{\partial t}\Big)u = 0\,, \tag{5.1.1}$$

where $v_j \in [-1, -\varepsilon]$. The corresponding reflection coefficient is found to be

$$r_d(\tau,\eta) = \prod_{j=1}^{d} \Big\{ \frac{\sqrt{1 - |\eta|^2/\tau^2} + v_j}{\sqrt{1 - |\eta|^2/\tau^2} - v_j} \Big\} . \tag{5.1.2}$$

5.1.1 PROPOSITION. *Let* $v_j \in [-1, -\varepsilon]$ . *Then* $Q_d$ *of Eq.(5.1.1) is strongly well-posed except for* $(\tau,\xi,\eta) = (0,0,0)$ . *Furthermore,* $r_d(\tau,\eta) < 1$ , *for all* $(\tau,\eta) \in \Gamma_\varepsilon$ *and*

$$\lim_{d\to\infty} r_d(\tau,\eta) = 0\,, \quad \textit{for } (\tau,\eta) \in \Gamma_\varepsilon .$$

PROOF: The statement regarding the reflection coefficient is obvious. We need to show that $Q_d$ is strongly well-posed. According to Thm.4.3.5, it is sufficient to study only one factor in Eq.(5.1.2). Thus we consider the mutual solution of

$$\tau^2 - (\xi^2 + |\eta|^2) = 0$$
$$\xi + v_j\tau = 0 .$$

The second equation implies that $\operatorname{Im}\xi$ and $\operatorname{Im}\tau$ have the same sign, showing that there are no eigenvalues. There are also no generalized eigenvalues except for $(\tau,\xi,\eta) = (0,0,0)$ . This completes the proof. ☐

As another example, we consider the Klein-Gordon equation

$$u_{tt} - \Delta u + u = 0 .$$

By substituting $i\tau$ for $u_t$ and so on, the dispersion relation is obtained; it is given by

$$\tau^2 - (\xi^2 + |\eta|^2 + 1) = 0\,.$$

It turns out that Eq.(5.1.1) is also a candidate for the $\mathcal{L}_\epsilon^\infty$ -best absorbing boundary condition of the Klein-Gordon equation. Furthermore, the previous proposition also holds.

Engquist and Majda [EnMa77] constructed a hierarchy of absorbing boundary conditions for the wave equation. In [EnMa77] and [EnMa79], they applied the theory of reflection of singularity [Nire73] and argued that a perfectly absorbing boundary condition is equivalent to requiring the solution to satisfy

$$\xi = \tau\sqrt{1 - \frac{|\eta|^2}{\tau^2}} \tag{5.1.3}$$

at the artificial boundary. They then approximated Eq.(5.1.3) by a Padé approximation to arrive at some local boundary conditions which are highly absorbing for high frequencies.

Higdon[Higd86a] later showed that the high-order Engquist-Majda boundary conditions are equivalent to

$$\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial t}\right)^d u = 0\,.$$

He further proved that, for certain approximations of the square root function in Eq.(5.1.3), the optimal absorbing boundary conditions of the wave equation must be expressible in the factorization form of Eq.(5.1.1), in which the parameter $v_j$ is interpreted as $|v_j| = \cos\theta_j$ , where $\theta_j$ is the angle of incidence of plane wave solutions. This result can be considered as a special case of our factorization theorem Thm.4.3.1.

## 5.2 Transonic Small Disturbance Equations

The basic small disturbance equation governing unsteady transonic flows in two dimensions is given by [CoCo86]

$$A\phi_{tt} + 2B\phi_{xt} = K^*\phi_{xx} + \phi_{yy} \tag{5.2.1}$$

where $K^* = K - (\gamma+1)\phi_x$ and $A = \frac{\kappa^2}{\delta^{2/3}}$ , $B = \frac{\kappa}{\delta^{2/3}}$ . In the above equation, $\phi$ is the disturbance velocity potential, $\kappa$ is the reduced frequency, $\delta, K, \gamma$ are constants. When the reduced frequency $\kappa$ is very small, $A \ll B$ . An approximation of Eq.(5.2.1), valid for low reduced frequencies, is obtained by letting $A = 0$ and $B = 1$ in Eq.(5.2.1).

In the far field, where $K^* = K - (\gamma + 1)\phi_x > 0$ , the equation is hyperbolic in the $t$ -direction. The equation is nonlinear, but the nonlinear effect is small in the farfield where the artificial boundaries are located. Therefore, in deriving absorbing boundary conditions for Eq.(5.2.1), we treat $K^*$ as a constant outside the computational domain and employ the theory developed in the previous chapter. This frozen coefficient technique is different from the method where the solution on the boundary is prescribed by the asymptotic solution of the equation. The former method still allows the solution to oscillate outside the computational domain but is governed by a simpler equation.

The artificial boundaries are illustrated in Fig.1.0.1. We will derive absorbing boundary conditions for each of the boundaries. The dispersion relation of Eq.(5.2.1) is given by

$$P(\tau, \xi, \eta) = A\tau^2 + 2B\tau\xi - K^*\xi^2 - \eta^2 = 0. \tag{5.2.2}$$

### 5.2.1 SIDEWALLS.

Working through the dispersion relation Eq.(5.2.2), we obtain

$$\frac{\eta^2}{(\tau + \frac{B}{A}\xi)^2} = \frac{A\eta^2}{\eta^2 + (K^* + \frac{B^2}{A})\xi^2} \le A\,.$$

From this, the $y$-component of group velocity can be found to satisfy $C_y \in [-\frac{1}{\sqrt{A}}, \frac{1}{\sqrt{A}}]$. For $(\tau, \xi) \in \mathbb{R}^2$, the equation Eq.(5.2.2) has two roots $\eta^+ = -\eta^-$, corresponding to two travelling modes and $0 < C_x(\eta^+) = -C_x(\eta^-)$. As a direct consequence of Thm.4.3.1, the $\mathcal{L}_\varepsilon^\infty$-best absorbing boundary condition must have the form

$$Q_d(x) = \prod_{j=1}^{d}(x - v_j), \tag{5.2.3}$$

where $v_j \in [-\frac{1}{\sqrt{A}}, -\varepsilon]$ for $y = -1$ and $v_j \in [\varepsilon, \frac{1}{\sqrt{A}}]$ for $y = 1$. Furthermore, $r_{Q_d} < 1$ by Thm.4.3.4.

In order to determine whether the boundary condition Eq.(5.2.3) is well-posed, we consider the common solution of

$$\eta + v_j(A\tau + B\xi) = 0$$

and Eq.(5.2.2). The above equation leads to $\frac{\operatorname{Im}\eta}{\operatorname{Im}\tau} = -v_j A > 0$. This relation shows that there are no eigenvalues. There are also no generalized eigenvalues except $(0, 0, 0)$. Thus the sidewall condition can be chosen as

$$\prod_{j=1}^{d}\{\frac{\partial}{\partial y} + v_j A\frac{\partial}{\partial t} + v_j B\frac{\partial}{\partial x}\}\phi = 0\,. \tag{5.2.4}$$

### 5.2.2 UPSTREAM AND DOWNSTREAM.

Solving for the group velocity from Eq.(5.2.2), we obtain

$$C_x = \frac{B}{A} - \frac{(K^* + \frac{B^2}{A})\xi}{A\tau + B\xi}$$

and

$$\frac{B}{A} - \sqrt{\frac{K^*}{A} + \left(\frac{B}{A}\right)^2} \le C_x \le \frac{B}{A} + \sqrt{\frac{K^*}{A} + \left(\frac{B}{A}\right)^2}.$$

The $\mathcal{L}_\varepsilon^\infty$ -best absorbing boundary condition of degree $d$ is

$$\prod_{j=1}^{d} \left\{ (B - v_j A)\frac{\partial}{\partial t} - (K^* + v_j B)\frac{\partial}{\partial x} \right\} \phi = 0 \tag{5.2.5}$$

where $v_j \in [\frac{B}{A} - \sqrt{\frac{K^*}{A} + \left(\frac{B}{A}\right)^2}, -\varepsilon]$ for the upstream boundary and $v_j \in [\varepsilon, \frac{B}{A} + \sqrt{\frac{K^*}{A} + \left(\frac{B}{A}\right)^2}]$ for the downstream boundary. It can be shown by the same method used earlier that the boundary condition Eq.(5.2.5) is strongly well-posed except for $(0, 0, 0)$ .

Engquist and Majda [EnMa80] derived absorbing boundary conditions for the transonic small disturbance equation of low reduced frequencies. They used an approach similar to but different from that of [EnMa77] discussed previously. Their derivation, however, did not provide a uniform approach in constructing the local boundary conditions. Different strategies had to be used to approximate the square root while handling the sidewall conditions. Furthermore, if a higher-order approximation is required, their method leads to two unpleasant difficulties, the determination of well-posedness and the selection of the parameters. Although general guidelines for selecting the parameters were given, these guidelines do not have a direct physical interpretation.

Absorbing boundary conditions for the low reduced frequency equation can be obtained by taking the limits $A \to 0$ and $B \to 1$ in Eq.(5.2.4) and Eq.(5.2.5). We now compare these results with the boundary conditions of [EnMa80].

### 5.2.3 SIDEWALLS.

Letting $A \to 0$ and $B = 1$ in Eq.(5.2.4), the condition becomes

$$\prod_{j=1}^{d} \left(\frac{\partial}{\partial y} + v_j \frac{\partial}{\partial x}\right)\phi = 0 . \tag{5.2.6}$$

The parameters $v_j$ must be in the interval $(-\infty, -\varepsilon]$ for $y = -1$ and in $[\varepsilon, \infty)$ for $y = 1$ . Even though the propagating speed in the $y$ -direction may be infinite, it can be justified to consider only the group velocities in a bounded interval, for the reasons given below.

The propagating speed of a plane wave in the $y$ -direction becomes infinite only if the wave is parallel to the $x$ -axis, i.e., $\xi = 0$ . In this case, it is easy to verify that the propagating speed downstream is also infinite and is much faster than the $y$ -component because

$$\frac{C_y}{C_x} \to 0 \quad \text{as} \quad \xi \to 0 .$$

This implies that the disturbance with infinite $y$ -direction speed will actually strike the downstream boundary first. Therefore in designing a sidewall boundary condition, we only have to consider those waves with finite propagating speed in the $y$ -direction. Thus, we may assume that there is a $\beta > 0$ such that

$$C_y \in [-\beta, \beta] .$$

The first-order boundary condition of Engquist and Majda is a special form of Eq.(5.2.6), with $v = v_1 = \sqrt{K^*}$ . This parameter yields a boundary condition that annihilates the waves travelling in the $y$ -direction only, i.e., $C_x = 0$ .

The second-order boundary condition of [EnMa80] is given by

$$\phi_{tx} + r_2 \phi_{xy} + r_1 \phi_{xt} = 0 . \tag{5.2.7}$$

where $r_1, r_2 > 0$ for the boundary $y = 1$ . Condition Eq.(5.2.7) is not a special case of the second-order condition obtained from Eq.(5.2.6), but rather that of the third-order condition of Eq.(5.2.6).

Taking three terms in Eq.(5.2.6) and expanding the factors, we have

$$\begin{aligned}\Big\{\frac{\partial^3}{\partial y^3}+&(v_1+v_2+v_3)\frac{\partial^2}{\partial y^2}\frac{\partial}{\partial x}\\ &+(v_1v_2+v_2v_3+v_3v_1)\frac{\partial}{\partial y}\frac{\partial^2}{\partial x^2}+v_1v_2v_3\frac{\partial^3}{\partial x^3}\Big\}\phi=0\,.\end{aligned}\tag{5.2.8}$$

We can then solve for $\phi_{yy}$ from the governing equation Eq.(5.2.1) (with $A=0$ ) to obtain

$$\phi_{yy}=2\phi_{xt}-K^*\phi_{xx}\,.\tag{5.2.9}$$

Differentiating Eq.(5.2.9) with respect to $y$ and inserting the result into Eq.(5.2.8), we obtain

$$2\phi_{xty}-K^*\phi_{xxy}+(v_1+v_2+v_3)\phi_{yyx}+(v_1v_2+v_2v_3+v_3v_1)\phi_{yxx}+v_1v_2v_3\phi_{xxx}=0\,.$$

When the above equation is integrated with respect to $x$ , the result is

$$2\phi_{ty}-K^*\phi_{xy}+(v_1+v_2+v_3)\phi_{yy}+(v_1v_2+v_2v_3+v_3v_1)\phi_{xy}+v_1v_2v_3\phi_{xx}=0\,.$$

Finally, substituting Eq.(5.2.9) into the above equation yields

$$\begin{aligned}2\phi_{ty}+(v_1v_2+v_2v_3+v_3v_1-K^*)\phi_{xy}+&2(v_1+v_2+v_3)\phi_{xt}\\ &+(v_1v_2v_3-K^*(v_1+v_2+v_3))\phi_{xx}=0\,.\end{aligned}$$

This condition is equivalent to Eq.(5.2.7) if $v_1,v_2,v_3$ are chosen such that

$$\begin{aligned}v_1v_2+v_2v_3+v_3v_1-K^*&=2r_2\\ v_1+v_2+v_3&=r_1\\ v_1v_2v_3-K^*(v_1+v_2+v_3)&=0\,.\end{aligned}$$

The above calculation reveals that any second-order condition in the form of Eq.(5.2.6) has less absorptivity than the second-order Engquist-Majda condition of the form Eq.(5.2.7) for some $r_1, r_2$ , and the latter in turn has less absorptivity than a third-order condition Eq.(5.2.6) for some choice of $v_1, v_2, v_3$ . Thus we may say that the second-order Engquist-Majda condition of the form Eq.(5.2.7) is a two-and-a-half-order condition of the form Eq.(5.2.6). However, since the parameters in Eq.(5.2.7) do not bear any immediate physical meaning, it is difficult to determine the optimal values of $r_1, r_2$ . The guidelines provided in [EnMa80] tend to annihilate the waves travelling at fast speed, which, as we discussed earlier, hit the downstream boundary first, hence has been absorbed by the downstream absorbing boundary condition. This argument is strongly supported by the results of our numerical experiments to be reported later.

It can also be shown [JiWo89] that the first- and the second-order upstream boundary conditions of [EnMa80] can be cast in the factorization form of Eq.(5.2.5) (with $A = 0$ and $B = 1$ ) but each parameter $v$ equals the same value $v = -\frac{K^*}{2}$ , which is tuned to the waves travelling most rapidly upstream. Thus the second-order upstream boundary condition of [EnMa80] is not optimal according to Thm.4.3.3.

Finally, we consider the downstream boundary conditions. By letting $A = 0$ and $B = 1$ in Eq.(5.2.5), we have

$$\prod_{j=1}^{d} \Big(\frac{\partial}{\partial t} - (K^* + v_j)\frac{\partial}{\partial x}\Big)\phi = 0\,, \tag{5.2.10}$$

where $v_j \in [\varepsilon, +\infty)$ . The first- and second-order conditions of [EnMa80] are equivalent to Eq.(5.2.10) with $v_1 = v_2 = +\infty$ , i.e.,

$$\phi_x = 0$$

for the first-order and

$$\phi_{xx} = 0 \tag{5.2.11}$$

for the second-order condition. Again, the second-order condition Eq.(5.2.11) is not optimal because it violates Thm.4.3.3.

## 5.3 Numerical Computations

In this section, we discuss the numerical implementation of the absorbing boundary conditions derived in the last section and present some computational results. For the purpose of illustrating the efficiency of these conditions, we compare the computational results obtained by using the boundary conditions of this thesis with those developed by Engquist and Majda in [EnMa80], both applied to the equation Eq.(5.2.1) for a constant $K^*$ in the region $\Omega = [-1, 1] \times [-1, 1]$. We note that the first-order conditions derived from the last section are the same as those derived in [EnMa80].

Specifically, we compare the following sets of conditions:

### 5.3.1 UPSTREAM CONDITIONS.

First-order:

$$\phi_t + v\phi_x = 0.$$

Second-order:

$$\phi_{xt} - \frac{2}{K^*}\phi_{tt} + \frac{1}{2}\phi_{yy} = 0, \qquad \text{(E \& M)}$$

$$\left(\frac{\partial}{\partial t} - (K^* + v_1)\frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial t} - (K^* + v_2)\frac{\partial}{\partial x}\right)\phi = 0. \tag{5.3.1}$$

### 5.3.2 DOWNSTREAM CONDITIONS.

First-order:

$$\phi_x = 0.$$

Second-order:

$$\phi_{xx} = 0, \qquad \text{(E \& M)}$$

$$\Big(\frac{1}{K^* + v}\frac{\partial}{\partial t} - \frac{\partial}{\partial x}\Big)\frac{\partial}{\partial x}\phi = 0. \qquad (5.3.2)$$

### 5.3.3 SIDEWALL $(y = 1)$ CONDITIONS.

First-order:

$$\phi_y + v\phi_x = 0.$$

Second-order:

$$\phi_{ty} + r_2\phi_{xy} + r_1\phi_{xt} = 0, \qquad \text{(E \& M)}$$

$$\Big(\frac{\partial}{\partial y} + v_1\frac{\partial}{\partial x}\Big)\Big(\frac{\partial}{\partial y} + v_2\frac{\partial}{\partial x}\Big)\phi = 0. \qquad (5.3.3)$$

### 5.3.4 PARAMETERS.

Theoretically, the parameters in the above boundary conditions should be determined by the minimization properties stated in the Def.4.2.3 and Def.4.2.4. Such a process, however, is neither necessary in practice, nor will it always yield the best result for all problems, because the definition of the $\mathcal{L}^\infty_\varepsilon$ - or $\mathcal{L}^2$ -best absorbing conditions itself depends on the arbitrary choice of $\varepsilon$ and the weight $\rho$ . In the present study, we use the strategies of § 4.4.

For the upstream conditions, the parameters must be in the interval $[-\frac{K^*}{2}, 0)$ . The first-order condition is tuned to the waves travelling most rapidly upstream by letting $v = -\frac{K^*}{2}$ . In the second-order condition Eq.(5.3.1), the two parameters are chosen as $v_1 = -\frac{K^*}{2}$ and $v_2 = -\frac{K^*}{4}$ , the mid-point of $[-\frac{K^*}{2}, 0)$ .

The parameters in the downstream conditions must be in $(0, +\infty)$ , which is unbounded. However, a transformation exists to map $(0, +\infty)$ to the bounded interval $(0, 1)$ , hence a method similar to the above can be used. Recall from § 5.2 that a factor of the reflection coefficient of the downstream condition is

$$\begin{aligned} R_j &= \frac{v_j - (K^* + v_j)\sqrt{1 - K^*(\frac{\eta}{\tau})^2}}{v_j + (K^* + v_j)\sqrt{1 - K^*(\frac{\eta}{\tau})^2}} \\ &= \frac{\frac{v_j}{(K^* + v_j)} - \sqrt{1 - K^*(\frac{\eta}{\tau})^2}}{\frac{v_j}{(K^* + v_j)} + \sqrt{1 - K^*(\frac{\eta}{\tau})^2}}, \end{aligned}$$

where $\frac{v_j}{K^* + v_j} \in (0, 1)$ .

Thus the first-order downstream condition is obtained by letting $\frac{v}{K^* + v} = 1$ , i.e., $v = +\infty$ . The remaining parameter $v$ in the second-order condition Eq.(5.3.2) is determined by requiring $\frac{v}{K^* + v} = \frac{1}{2}$ , therefore, $v = K^*$ .

In the case of the sidewalls, we have to deal with the genuine unbounded interval. For reasons mentioned in § 5.2, $v = \sqrt{K^*}$ in the first-order sidewall ( $y = 1$ ) condition. Parameters $v_1$, $v_2$ in Eq.(5.3.3) are determined by minimizing the integral

$$\int_0^\beta e^{-\alpha(x - \sqrt{K^*})^2} R_2(x)^2 dx,$$

where, by § 4.4, $R_2 = (\frac{x - v_1}{x + v_1})(\frac{x - v_2}{x + v_2})$ .

By a change of variables, the above integral becomes

$$\sqrt{K^*} \int_0^{\tilde{\beta}} e^{-\tilde{\alpha}(x-1)^2} (\frac{x - \tilde{v}_1}{x + \tilde{v}_1})^2 (\frac{x - \tilde{v}_2}{x + \tilde{v}_2})^2 dx,$$

where $\tilde{\beta} = \frac{\beta}{\sqrt{K^*}}$, $\tilde{\alpha} = \alpha K^*$ and $\tilde{v}_j = \frac{v_j}{\sqrt{K^*}}$ . $\tilde{v}_1$ and $\tilde{v}_2$ can be computed by a numerical integration and optimization algorithm. For $\tilde{\beta} = 10.0$ and $\tilde{\alpha} = 0.1$ , we found, correct to the first digit, $\tilde{v}_1 = 3.0$ and $\tilde{v}_2 = 0.4$, hence $v_1 = 3.0\sqrt{K^*}$ and $v_2 = 0.4\sqrt{K^*}$ in Eq.(5.3.3). This particular choice of parameters yields a reflection coefficient $R_2 \leq 0.05$ for all wave modes with group velocities in the range $[0.25\sqrt{K^*},\ 5.0\sqrt{K^*}]$.

Finally, in this study, $r_1 = 6.4641\sqrt{K^*}$ and $r_2 = 5.4641K^*$ in the (E & M) second-order condition, according to one of the guidelines of [EnMa80].

### 5.3.5 DISCRETIZATION.

Eq.(5.2.1) is discretized by a semi-implicit scheme described in [BaGo77], where the $x$-direction is treated explicitly and the $y$-direction is kept implicit. The resulting difference equation is solved by marching both in the $x$- and $t$-direction. At a given time level, tridiagonal equations are solved along the $y$-direction, successively marching in the downstream direction. For stability, the *C.F.L.* condition $K^*\dfrac{\Delta t}{\Delta x} \leq 2$ must be satisfied. The scheme is nondissipative when $K^*\dfrac{\Delta t}{\Delta x} = 2$.

For the first-order boundary conditions and the second-order Engquist-Majda (E & M) boundary conditions, we use exactly the discretization suggested in [EnMa80]. Since each factor in Eq.(5.3.1) and Eq.(5.3.3) is in the form of the first-order operator of the corresponding first-order boundary condition, the discretizations of Eq.(5.3.1) and Eq.(5.3.3) are obtained by applying twice the difference operators which discretize the corresponding first-order boundary conditions of Engquist and Majda. For example, in [EnMa80], the boundary condition

$$\phi_y + v\phi_x = 0,$$

is discretized by

$$\left\{\frac{I-E_y^{-1}}{\Delta y}(I+E_x^{-1})+v\frac{I-E_x^{-1}}{\Delta x}(I+E_y^{-1})\right\}\phi^{n+1}_{j,m+1}=0,$$

where $E_x$, $E_y$ are shift operators in $x$, $y$ respectively, $E_x\phi^n_{j,m}=\phi^n_{j+1,m}$, $E_y\phi^n_{j,m}=\phi^n_{j,m+1}$. By applying this twice, the difference equation for Eq.(5.3.3) becomes

$$\begin{aligned}&\left\{\frac{I-E_y^{-1}}{\Delta y}(I+E_x^{-1})+v_1\frac{I-E_x^{-1}}{\Delta x}(I+E_y^{-1})\right\}\\ &\qquad\times\left\{\frac{I-E_y^{-1}}{\Delta y}(I+E_x^{-1})+v_2\frac{I-E_x^{-1}}{\Delta x}(I+E_y^{-1})\right\}\phi^{n+1}_{j,m+1}=0. \qquad (5.3.4)\end{aligned}$$

Condition Eq.(5.3.2) is discretized by an explicit scheme, backward in $x$.

The stability of these difference schemes can be determined by the GKS theory [GKS72] (see 6.1.2), or by numerical experiments. Because of the factorization form, the difference schemes for the higher-order boundary conditions are stable if and only if, by the GKS theory, the schemes for the corresponding first-order conditions are stable. The discretization of the first-order boundary conditions is discussed in detail in [EnMa80], and although no analysis has been attempted, these schemes are found to be stable by numerical experiments, even for the nonlinear problem.

Since the interior equation is solved line by line by marching in the downstream direction, the boundary condition Eq.(5.3.4) which requires values of $\phi$ at $j-1$, $j-2$ cannot be applied to the immediate neighboring point of the corner of the upstream and the sidewall boundaries. This difficulty is overcome by using the following condition at the upstream corner:

$$\alpha(x)\Big(\frac{\partial}{\partial y}+v\frac{\partial}{\partial x}\Big)\phi+(1-\alpha(x))\Big(\frac{\partial}{\partial y}+v_1\frac{\partial}{\partial x}\Big)\Big(\frac{\partial}{\partial y}+v_2\frac{\partial}{\partial x}\Big)\phi=0,$$

where $0\le\alpha\le 1$. $\alpha$ decreases smoothly from $\alpha=1$ at the upstream boundary to $\alpha=0$. In our study, the transition region of $\alpha$ from 1 to 0 extends to 2 grid points.

There is no difficulty at the downstream corners, since these corner points are not involved in the computation.

### 5.4.6 THE ENERGY.

In the following computations, we will use the energy of solutions as a means of measurement. The energy

$$E(n) = \sum_{j,\ m} K^{*} \left\{ \frac{\phi^n_{j+1,m+1} + \phi^n_{j+1,m} - \phi^n_{j,m+1} - \phi^n_{j,m}}{2\Delta x} \right\}^2 \Delta x \Delta y + \sum_{j,\ m} \left\{ \frac{\phi^n_{j+1,m+1} + \phi^n_{j,m+1} - \phi^n_{j+1,m} - \phi^n_{j,m}}{2\Delta y} \right\}^2 \Delta x \Delta y, \tag{5.3.5}$$

is the discrete version of

$$E(t) = \iint_{\substack{|x|\le 1 \\ |y|\le 1}} \left( K^{*}\ \phi_x^2 + \phi_y^2 \right) dx dy.$$

It is easy to verify that if $\phi$ satisfies Eq.(5.2.1) in $|x| < 1$ and $|y| < 1$, and the perfectly reflecting boundary conditions

$$\phi = 0, \quad \text{at} \quad x = \pm 1 \text{ and } y = \pm 1, \tag{5.3.6}$$

then the energy of solutions is conserved, that is,

$$\frac{d}{dt} E(t) \equiv 0, \quad \text{for all } t. \tag{5.3.7}$$

In order to examine the effects of each absorbing boundary condition, without being influenced by the presence of the other boundary conditions, we will use a strategy where the absorbing boundary condition will be prescribed only at one boundary in each experiment, and the perfectly reflecting boundary condition

Eq.(5.3.6) will be imposed on the remaining boundaries. In the first set of calculations, the energy Eq.(5.3.5) of solutions in the domain $\Omega = [-1,1] \times [-1,1]$ will be computed. Therefore, the rate of decrease of the energy will demonstrate the ability of the test boundary condition to radiate the energy away from the computational domain, because by Eq.(5.3.7), the energy of the solution will be totally reflected back from the other three perfectly reflecting boundaries.

In all computations, $K^* = 1$ is used in the equation Eq.(5.2.1), and the region $\Omega = [-1,1] \times [-1,1]$ is covered by a $60 \times 40$ uniform mesh, with a $CFL$ condition of 1.997 . For an initial value, we use a pulse of compact support, a piece of the radially symmetrical sine function. The energy of solutions is calculated after every 10 time steps, with a total computation of 230 time steps. The initial pulse will soon spread out in a parabolic wavefront [EnMa80], and total reflections will occur at the three perfectly reflecting boundaries, so after some time, the disturbance that strikes the test boundary will consist of wave packets with a fairly large spectrum of frequencies and wave numbers.

Fig.5.3.1 shows the energy $E^0(n)$ , given by Eq.(5.3.5), of the solution of Eq.(5.2.1) with the perfectly reflecting boundary conditions Eq.(5.3.6) at different time steps. The decrease in the energy is due to the dissipation of the difference scheme. In order to eliminate the effect of this dissipation in our study, the energy of the solutions for other calculations will all be scaled by $E^0(n)$ , i.e., the energy

$$e(n) = \left(\frac{E(n)}{E^0(n)}\right) \times 100,$$

will be used for other calculations. Therefore, $e(n) \equiv 100$ if the perfectly reflecting boundary condition $\phi = 0$ is prescribed at all boundaries.

The results of the calculations involving boundary conditions on the upstream, downstream and sidewall ( $y = 1$ ) boundaries are presented in Fig.5.3.2, Fig.5.3.3

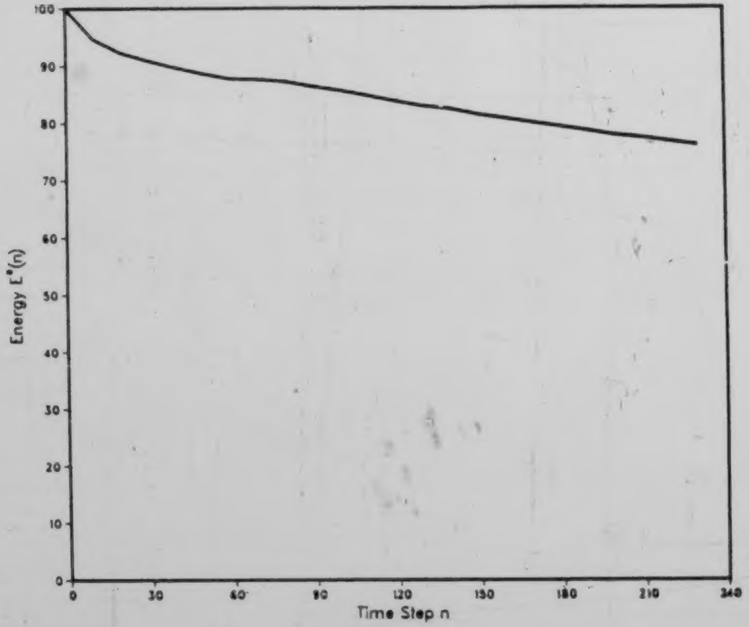

**Figure 5.3.1** The energy of solutions
Perfectly reflecting boundary conditions are used on all four sides.
The decrease in the energy is due to the dissipation of the scheme.

and Fig.5.3.4 respectively. The curves of the energy $e(n)$ all follow one pattern: during the short period of the initial time (about 30 time steps), the three (one first-order and two second-order) boundary conditions produce roughly the same results, but they differ significantly after a long time period (after about 130 time steps).

In the second set of calculations, we measure the energy of the reflected waves of different absorbing boundary conditions. The reflected wave is obtained by comparing the solution of an absorbing boundary condition with the free-space solution which is calculated in a larger computational domain. For example, in the computations involving upstream boundary conditions, the free-space solution is calculated in the region $\Omega \cup \bar{\Omega}$ as illustrated in Fig.5.3.5. Then an absorbing boundary condition is used on the boundary $x = -1$, and the corresponding

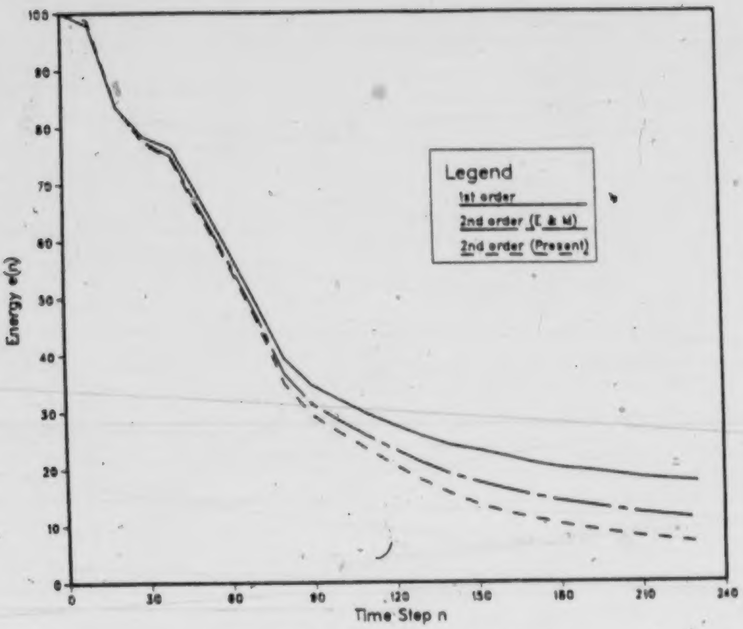

**Figure 5.3.2** The energy of solutions

Absorbing boundary conditions are used on upstream boundary. $\upsilon_1 = -\frac{K^*}{2}, \upsilon_2 = -\frac{K^*}{4}$ in condition Eq.(5.3.1)

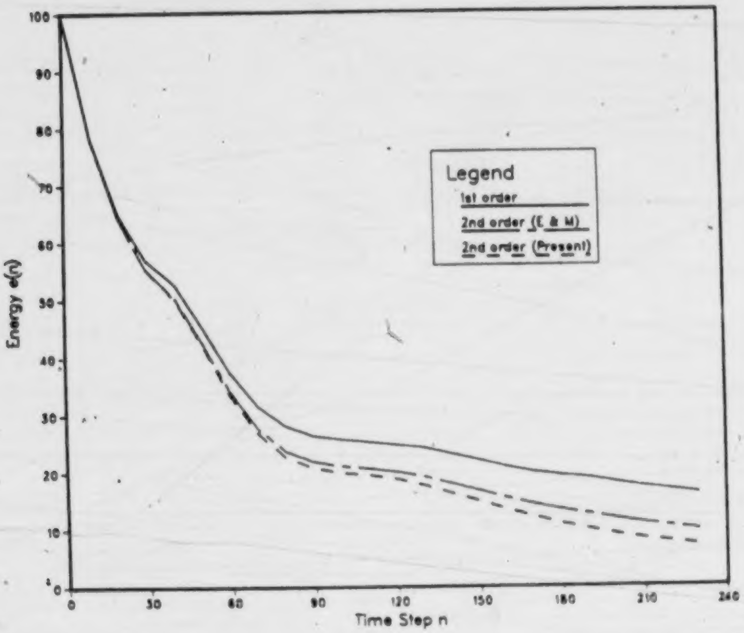

**Figure 5.3.3** The energy of solutions

Absorbing boundary conditions are used on downstream boundary. $\upsilon = K^*$ in condition Eq.(5.3.2)

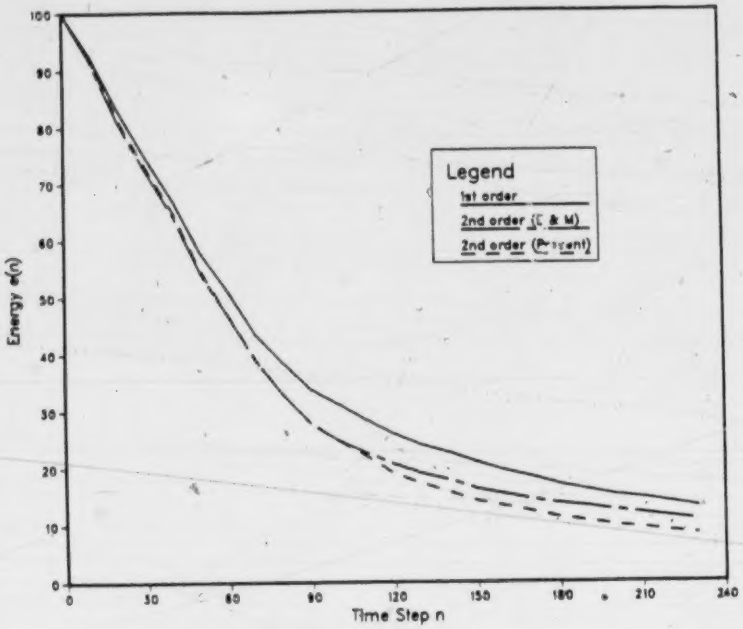

**Figure 5.3.4** The energy of solutions

Absorbing boundary conditions are used on sidewall( $y = 1$ ). $v_1 = 3.0\sqrt{K^*}, v_2 = 0.4\sqrt{K^*}$ in condition Eq.(5.3.3). $r_1 = 6.4641\sqrt{K^*}, r_2 = 5.4541K^*$ in (E&M).

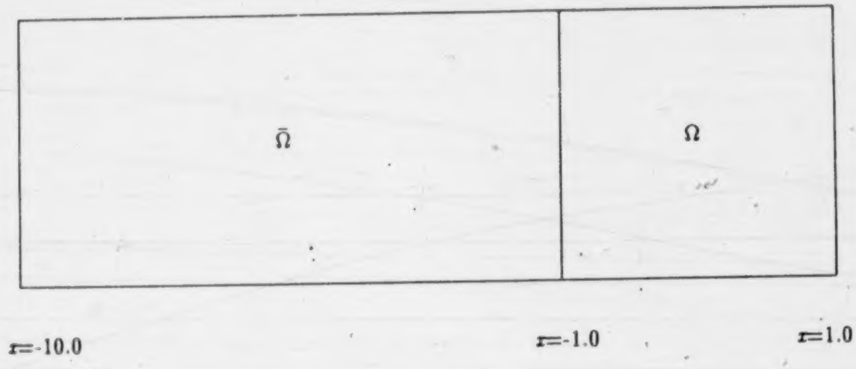

**Figure 5.3.5** Larger computational domain

solution is computed in the region $\Omega$ . The difference of these two solutions in $\Omega$ can be considered as the reflection caused by the absorbing boundary condition. Similar methods are used for the downstream and the sidewall boundaries.

The results for the energy of reflected waves are shown in Fig.5.3.6, Fig.5.3.7 and Fig.5.3.8. As before, the energy is calculated at every 10 time steps. In each graph, the curves are scaled so that the maximum energy in the first-order boundary condition is 100 .

These numerical results clearly show the improved performance of the second-order absorbing boundary conditions over the first-order conditions. In long time computations, the second-order conditions Eq.(5.3.1), Eq.(5.3.2) and Eq.(5.3.3) produce less reflections than the corresponding (E & M) conditions. However, the second-order (E & M) conditions generate smaller reflections during the very short period of the initial time because they are tuned to the waves travelling most rapidly. These observations agree with the analysis of the reflection coefficients given in § 5.2.

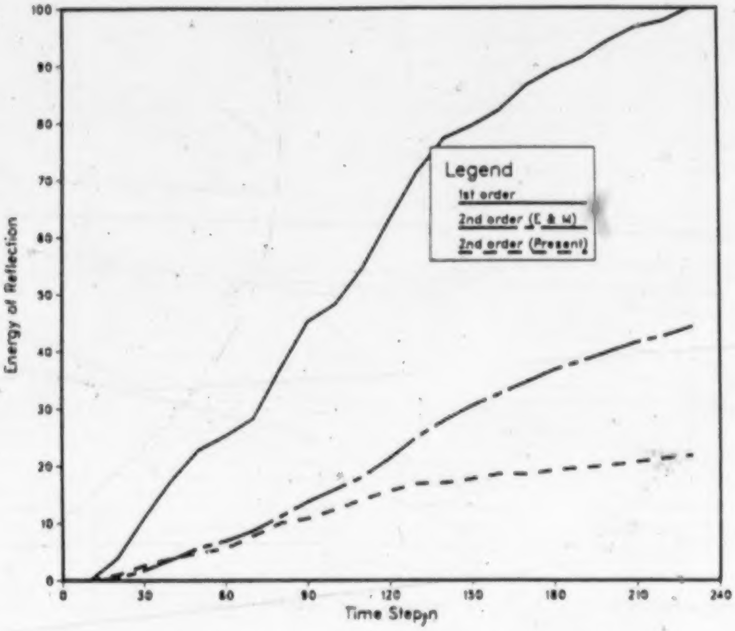

**Figure 5.3.6** The energy of reflected waves: Upstream.

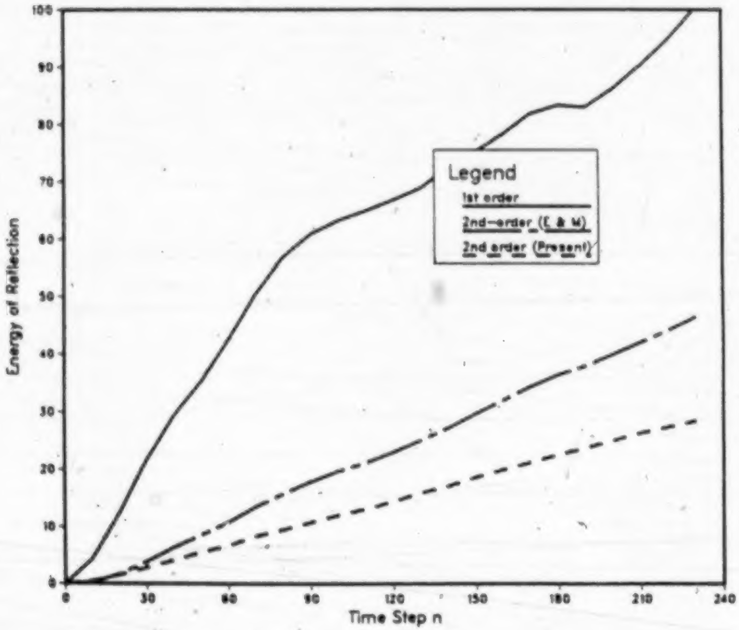

**Figure 5.3.7** The energy of reflected waves: Downstream.

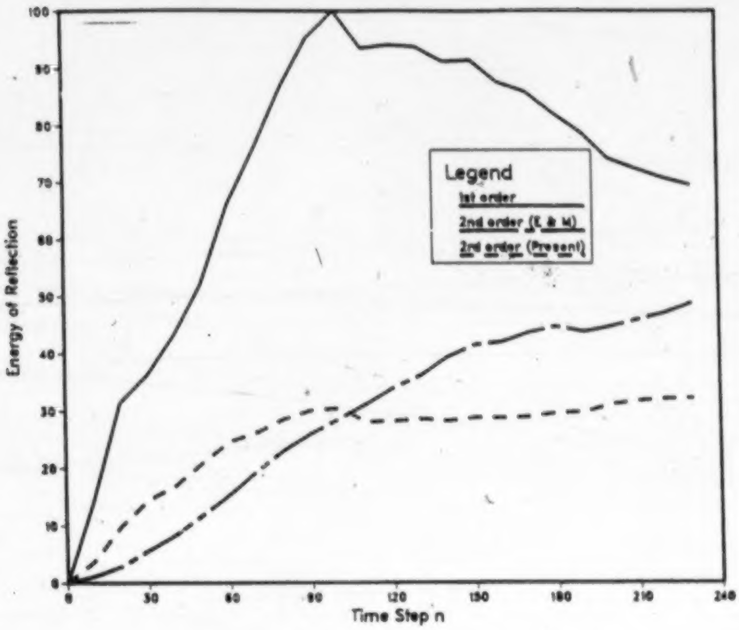

**Figure 5.3.8** The energy of reflected waves: Sidewall( $y = 1$ ).

# CHAPTER 6

# ABSORBING BOUNDARY CONDITIONS FOR FINITE DIFFERENCE SCHEMES

In Chapter 4, we developed a class of absorbing boundary conditions for hyperbolic differential equations. This technique can also be applied to difference equations resulting from the discretization of hyperbolic equations.

A usual practice to derive absorbing boundary conditions for the approximating difference equation is to determine an absorbing boundary condition for the original differential equation first and then to apply the discretizations of the condition to the difference equation. While this approach works well for many problems, there are circumstances where it is better to design a discrete absorbing boundary condition directly for the difference equation.

One such case is where solutions essentially contain very high frequencies that are described with only a few points per wave length. When this happens, the structure of the discrete solutions described by the differential equation may be quite different from that of its continuous ones.

Another example is a differential equation that needs no artificial boundary conditions, but its discretization does. In this case the absorbing boundary condition has to be constructed directly from the difference equation because there is no continuous counterpart to work with. Let us consider the following classical example.

$$\begin{aligned} u_t - u_x &= 0\,, \quad 0 < x < a, t \geq 0 \\ u &= g\,, \quad x = a, t > 0 \\ u &= f\,, \quad 0 \leq x \leq a, t = 0\,. \end{aligned} \tag{6.0.1}$$

No boundary condition should be given at the open boundary $x = 0$ because the problem as it stands is well-posed.

Suppose Eq.(6.0.1) is approximated by a leap-frog (LF) scheme. The resulting difference equation is

$$u_j^{n+1} = u_j^{n-1} + \lambda(u_{j+1}^n - u_{j-1}^n), \quad \lambda = \frac{\Delta t}{\Delta x}, \quad 0 \leq j \leq j_a. \qquad (6.0.2)$$

The solution of Eq.(6.0.2) can be obtained by marching in time $n$, and at each time step, sweeping downstream (assume $u_j^1$ is approximated by some first-order accurate scheme). A difficulty arises because $u_0^2$ cannot be determined. Consequently, at the subsequent time levels, $u_1^3, u_2^4, \ldots$ cannot be determined. For the LF scheme to be solvable, either the computational domain must be extended beyond $x = 0$, which is not recommended in consideration of computational cost, or an additional boundary condition must prescribed at $x = 0$. If the latter approach is chosen, the boundary condition cannot be arbitrary, because otherwise one may obtain a different solution. However, the correct boundary condition has to come from the difference equation, not the differential equation, because it is the discretization that needs such an extra boundary condition.

Fortunately, most of the theory developed in previous chapters is applicable to difference equations as well. The theory of stability for discrete initial boundary value problems is quite analogous to the theory of well-posedness for continuous IBVPs. The stability and its connection with wave propagation have been studied extensively by Trefethen [Tref82]. In the next section, we will review some results from [Tref82]. In section two, the technique of Chapter 4 will be applied to finite difference schemes to construct absorbing boundary conditions. In section three, we present some examples for one-dimensional problems.

## 6.1 General Considerations

Assume that a difference scheme of some hyperbolic equation is given by

$$P(E_t, E_x, E_y)u = 0\,, \quad n > 0, 0 < j < j_a, m \in \mathbb{Z} \tag{6.1.1a}$$

$$B(E_t, E_x, E_y)u = 0\,, \quad n > 0, j = 0 \tag{6.1.1b}$$

$$A(E_t, E_x, E_y)u = g\,, \quad n > 0, j = j_a \tag{6.1.1c}$$

$$u = 0\,, \quad n \le 0\,. \tag{6.1.1d}$$

In Eq.(6.1.1), $E_t, E_x, E_y$ denote the shift operators in $t, x, y$, respectively, i.e.,

$$E_t u^n_{j,m} = u^{n+1}_{j,m}\,, \quad E_x u^n_{j,m} = u^n_{j+1,m}\,, \quad E_y u^n_{j,m} = u^n_{j,m+1}\,.$$

If a plane wave solution

$$e^{i(\tau \Delta t n + \xi \Delta x j + \eta \Delta y m)} \tag{6.1.2}$$

is inserted into Eq.(6.1.1a), the result is the dispersion relation

$$P(z, \kappa, e^{i\eta\Delta y}) = 0\,, \tag{6.1.3}$$

where $z = e^{i\tau\Delta t}, \kappa = e^{i\xi\Delta x}$. The von Neumann condition for stability of the initial value problem Eq.(6.1.1a) and Eq.(6.1.1d) requires that Eq.(6.1.3) has no roots $z$, $|z| > 1$ for all $(\kappa, \eta) \in \mathbb{C} \times \mathbb{R}$ with $|\kappa| = 1$.

6.1.1 Assumption. *We assume that $P$ satisfies the von Neumann stability condition, i.e.,*

$$P(z, \kappa, e^{i\eta\Delta y}) \ne 0$$

*for all $(z, \kappa, \eta) \in \mathbb{C} \times \mathbb{C} \times \mathbb{R}$ such that $|z| > 1$ and $|\kappa| = 1$.*

Under this assumption, for any $(z, \eta) \in \mathbb{C} \times \mathbb{R}, |z| > 1$, a root $\kappa$ of Eq.(6.1.3) satisfies $|\kappa| \ne 1$. Thus we denote by $\kappa_+$ those roots for which $|\kappa_+| < 1$ and the rest of the roots are denoted by $\kappa_-$.

Associated with each wave solution Eq.(6.1.2) is a speed of translation defined by [Tref82] (also see discussion in § 3.2)

$$\mathring{C}_x = -\frac{1}{\lambda}\frac{\log|z|}{\log|\kappa|}.$$

Therefore a plane wave solution Eq.(6.1.2) is said to be leftgoing if $\mathring{C}_x \le 0$ and rightgoing if $\mathring{C}_x \ge 0$ . Clearly, the mode $(z, \kappa_+, \eta)$ is rightgoing and $(z, \kappa_-, \eta)$ leftgoing.

A necessary and sufficient condition for the stability of the initial boundary value problem Eq.(6.1.1) has been developed by Gustafsson *et al.* [GKS72]; it is an analogue of the UKC condition.

6.1.2 GKS STABILITY CRITERION. *The boundary condition Eq.(6.1.1b) is stable if and only if Eq.(6.1.3) and* $B(z, \kappa, e^{i\eta\Delta y}) = 0$ *have no common solution* $(z, \kappa, \eta) \in \mathbb{C} \times \mathbb{C} \times \mathbb{R}$ *for which* $|z| \ge 1$ *and* $\mathring{C}_x(z, \kappa, \eta) \ge 0$ .

If the above criterion is violated by $(z, \kappa, \eta)$ , then we say $(z, \kappa, \eta)$ is an *eigenvalue* if $|z| \ge 1, |\kappa| < 1$ ; it is a *generalized eigenvalue* if $|z| = |\kappa| = 1$ . In case $|z| = |\kappa| = 1$ , $\mathring{C}_x$ is defined by the limit

$$\lim_{|z|\to 1} \mathring{C}_x(z, \kappa, \eta) = C_x(z, \kappa, \eta),$$

where $C_x = -\dfrac{\partial \tau}{\partial \xi}$ is the group velocity of a wave packet of the form Eq.(6.1.2).

As an example, we consider the LF scheme Eq.(6.0.2). When $u = z^n \kappa^j$ is inserted into Eq.(6.0.2), the result is

$$\sin \tau\Delta t = \lambda \sin \xi \Delta x \,, \quad \lambda = \frac{\Delta t}{\Delta x}. \tag{6.1.4}$$

The relation between $\tau\Delta t$ and $\xi\Delta x$ is illustrated in Fig.6.1.1. Let us compare this with the dispersion relation of the differential equation Eq.(6.0.1), which is a

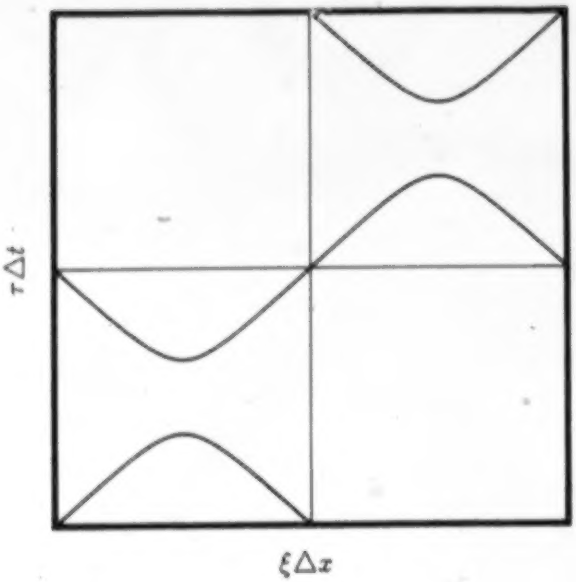

**Figure 6.1.1** Dispersion relation of LF, $\lambda < 1$

straight line of 45 degree in the $\xi - \tau$ plane. The former is a good approximation of the latter only in the vicinity of $(\tau\Delta t, \xi\Delta x) = (0,0)$. This means that a solution of Eq.(6.0.1) can be well approximated by a discrete solution only when the former can be well resolved on the mesh. That is, $\Delta t$, $\Delta x$ are small enough so that $(\tau\Delta t, \xi\Delta x) \approx (0,0)$ for all modes $(\tau, \xi)$ of which the solution is composed.

From Fig.6.1.1, it is clear that there is a value $\tau_c$, *cutoff frequency*, such that for each $\tau \in [0, \frac{\tau_c}{\Delta t}]$, the dispersion relation Eq.(6.1.4) has two roots $\xi^+, \xi^- \in [0, \frac{\pi}{\Delta x}]$. They correspond to two travelling modes, the rightgoing $(z, \kappa_+)$ and the leftgoing $(z, \kappa_-)$. It is this rightgoing mode that is responsible for the requirement of an additional boundary condition at $x = 0$.

The rightgoing waves corresponding to $(z, \kappa_+)$ are nonphysical; they are totally supported by the discretization. They are propagated not only at the

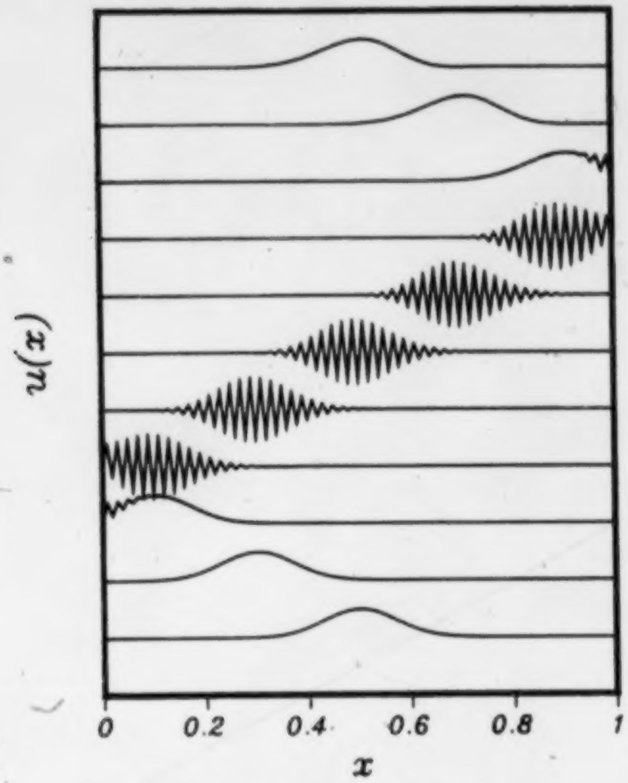

**Figure 6.1.2** Parasitic Waves

The LF scheme is solved in $[0,1]$ . Boundary condition $u^n = 0$ is used at both boundaries. The initial wave packet travels at the correct group velocity $-1$ ; upon reaching the boundary $x = 0$ , it is reflected as a parasitic wave travelling at a wrong speed. The graph shows the solution of Eq.(6.0.2) at different time steps.

wrong speed, but even in the wrong direction! Such waves are referred to as *parasitic waves*. Parasitic waves exist regardless of the mesh size. In general, they contain high frequencies that are described with only a few points per wave length. In Fig.6.1.1, parasitic waves are those for which $\xi \Delta x \in [\frac{\pi}{2}, \pi]$.. These waves have a spatially sawtoothed shape as illustrated in Fig.6.1.2.

## 6.2 Absorbing Boundary Conditions

In order to eliminate reflections from an artificial boundary, we apply the nonreflecting principle. As we have shown in Chapter 4, this amounts to requiring that

$$B(z,\kappa,e^{i\eta\Delta y}) = |\mathring{C}_x(z,\kappa,e^{i\eta\Delta y})| + \mathring{C}_x(z,\kappa,e^{i\eta\Delta y}) = 0 \tag{6.2.1}$$

be satisfied at the boundary $x = 0$ .

Eq.(6.2.1) is the symbol of a difference operator defined through a discrete Fourier transform. The definition and properties of the discrete Fourier transforms can be found in [ViBo82]. It can be verified, by a Fourier transform process similar to that of § 4.1, that Eq.(6.2.1) defines a perfectly absorbing boundary condition. The details will be omitted. The argument in this chapter will simply be based on reflection coefficients. Thus we regard a boundary condition as a good absorbing condition if it produces a small reflection coefficient.

For each $(z,\eta) \in \mathbb{C} \times \mathbb{R}$, $|z| \geq 1$ , let $\kappa_+, \kappa_-$ be the roots of the dispersion relation Eq.(6.1.3), corresponding to the rightgoing and leftgoing wave solutions respectively. Then the general solution of Eq.(6.1.1a) is

$$u^n_{j,m} = c_- z^n \kappa_-^j e^{i\eta\Delta y m} + c_+ z^n \kappa_+^j e^{i\eta\Delta y m} . \tag{6.2.2}$$

Thus the reflection coefficient of boundary condition Eq.(6.1.1b) can be obtained by substituting Eq.(6.2.2) into Eq.(6.1.1b) and solving for $c_+$ in terms of $c_-$ . That is,

$$r_B(z,\kappa,e^{i\eta\Delta y}) = \left| \frac{B(z,\kappa_-,e^{i\eta\Delta y})}{B(z,\kappa_+,e^{i\eta\Delta y})} \right| . \tag{6.2.3}$$

When the boundary condition Eq.(6.2.1) is inserted into Eq.(6.2.3), the result is $r_B = 0$ , because $\mathring{C}_x(z,\kappa_-,e^{i\eta\Delta y}) \leq 0$ , implying $B(z,\kappa_-,e^{i\eta\Delta y}) = 0$ . This verifies that Eq.(6.2.1) is a perfectly absorbing boundary condition. In the absence of evanescent waves, i.e., when the solution consisting only of the modes $(z,\kappa,\eta)$ with $|z| = |\kappa| = 1$ , Eq.(6.2.1) is equivalent to

$$B(\kappa,\eta) = |C_x(z,\kappa,e^{i\eta\Delta y})| + C_x(z,\kappa,e^{i\eta\Delta y}) = 0 , \tag{6.2.4}$$

where $C_x$ is the $x$-component of the group velocity given by

$$C_x = -\frac{\partial \tau}{\partial \xi} = \frac{P_\xi}{P_\tau}.$$

In order to construct an absorbing boundary condition Eq.(6.2.4) explicitly, we need to find $P_\xi$ and $P_\tau$. Because $P = P(z, \kappa, e^{i\eta\Delta y})$ where $z = e^{i\tau\Delta t}$ and $\kappa = e^{i\xi\Delta x}$, a direct calculation shows that

$$P_\tau(z, \kappa, e^{i\eta\Delta y}) = i\Delta t z P_z(z, \kappa, e^{i\eta\Delta y}), \quad P_\xi(z, \kappa, e^{i\eta\Delta y}) = i\Delta x \kappa P_\kappa(z, \kappa, e^{i\eta\Delta y}).$$

However, Eq.(6.2.4) is a nonlocal condition which is not numerically useful. To obtain its local approximation, we employ the same technique as in § 4.2 to arrive at some high-order local absorbing boundary conditions. The results can be summarized in the following.

6.2.1 THEOREM. *The GKS stable,* $\mathcal{L}^\infty_\epsilon$ *-best absorbing boundary condition for Eq.(6.1.1) has the form*

$$B_d(E_t, E_x, E_y)u^n_{j,m} =$$
$$\prod_{h=1}^{d} \Big(P_\kappa(E_t, E_x, E_y)E_x - v_h \lambda P_z(E_t, E_x, E_y)E_t\Big)u^n_{j,m} = 0, \ \lambda = \frac{\Delta t}{\Delta x}, \tag{6.2.5}$$

*where* $v_h \in [\alpha, -\varepsilon]$ *and* $\alpha = \min C_x(z, \kappa, e^{i\eta\Delta y})$. *Furthermore,*

$$v_h \neq v_k, \quad \text{if} \quad h \neq k.$$

Note that each factor in Eq.(6.2.5) annihilates wave packets travelling at a group velocity of $v_h$.

The implementation of the above boundary condition is not obvious; it may not be directly applied to the grid solution $u^n_0$ because in general $P$, hence

$B$ , contains a term $E_x^{-1}$ . The rest of this section is devoted to this issue. For definiteness of discussion, some assumptions must be made.

We first suppose that exactly one boundary condition is needed at each boundary $x = 0$ and $x = a > 0$ for Eq.(6.1.1) to be solvable. A typical scheme for a second-order partial differential equation is the three point multilevel scheme, which can be written in the form

$$P(E_t, E_x) = E_t\{a_+^1 E_x + a_o^1 + a_-^1 E_x^{-1}\} + \sum_{h=-r+2}^{0} E_t^h \{a_+^h E_x + a_o^h + a_-^h E_x^{-1}\} . \tag{6.2.6}$$

The solution of this scheme can be obtained by marching in time. When the scheme is implicit, i.e., if $a_+^1$ or $a_-^1 \neq 0$ , a tridiagonal matrix inversion is needed at each time level. Eq.(6.2.6) is explicit if $a_+^1 = a_-^1 = 0$ ; no matrix inversion is needed in this case. The leap-frog scheme is an example of Eq.(6.2.6), where

$$P_{LF}(E_t, E_x) = E_t - \lambda(E_x - E_x^{-1}) - E_t^{-1} .$$

For multidimensional problems, the coefficients $a$ in Eq.(6.2.6) are functions of $E_y, E_y^{-1}$ .

Solving for $P_z$ and $P_\kappa$ from Eq.(6.2.6) and substituting them into Eq.(6.2.5), one derives an absorbing boundary condition of degree $d$ . We consider one factor first, i.e., the absorbing boundary condition of degree 1 . The symbols of both the absorbing boundary condition of degree 1 and the interior scheme Eq.(6.2.6) can be expressed as

$$B_1(z, \kappa) = q_1(z, z^{-1})\kappa^{-1} + q_2(z, z^{-1}, \kappa) = 0 \tag{6.2.7a}$$

$$P(z, \kappa) = p_1(z, z^{-1})\kappa^{-1} + p_2(z, z^{-1}, \kappa) = 0 , \tag{6.2.7b}$$

where $p, q$ are some polynomials. By eliminating $\kappa^{-1}$ from Eq.(6.2.7), an equivalent boundary condition which we still denote by $B_1$ is obtained:

$$B_1(z,\kappa) = z\{b^1_+\kappa + b^1_o\} + \sum_{h=-r+2}^{0} z^h\{b^h_+\kappa + b^h_o\}. \tag{6.2.8}$$

Applying Eq.(6.2.8) to $u^n_0$, we obtain a complete tridiagonal system for $u^{n+1}_j, j = 0, 1, \dots, j_a$. If the interior scheme Eq.(6.2.6) is explicit, a diagonal system results, so that no matrix inversion is necessary.

Absorbing boundary conditions of high degree can be obtained by repeated application of Eq.(6.2.8). In general,

$$B_d(z,\kappa) = z^{-d+1} \prod_{h=1}^{d} B_h\,, \tag{6.2.9}$$

where each $B_h$ is of form Eq.(6.2.8) with some parameter $v_h$. Application of Eq.(6.2.9) to $u^n_0$ yields a complete system for $u^{n+1}_j, j = 0, 1, \dots, j_a$. In general, the system is no longer tridiagonal; however, it is still diagonal if Eq.(6.2.6) is explicit.

## 6.3 Examples

In this section, we provide two concrete examples to illustrate how the absorbing boundary conditions can be derived from the discrete equations. We also present some numerical results to support the discussion given earlier.

### 6.3.1 THE LEAP-FROG SCHEME.

We first consider the leap-frog scheme Eq.(6.0.2). It follows from the previous section that the necessary form of a stable $\mathcal{L}^\infty_\varepsilon$-best absorbing

$$B_d(E_t, E_x)u^n_j = \prod_{h=1}^{d} \left\{(E_x + E_x^{-1}) + v_h(E_t + E_t^{-1})\right\}u^n_j = 0\,, \tag{6.3.1}$$

where $v_h \in [-1, -\varepsilon]$ .

We need to determine the stability of the boundary condition Eq.(6.3.1). It suffices to show that Eq.(6.3.1) admits no eigenvalues because any boundary condition of the form Eq.(6.2.5) admits no generalized eigenvalues except for $z = \kappa = 1$ (which is also unavoidable). Thus we show that the following equations have no common solution $(z, \kappa) \in \mathbb{C} \times \mathbb{C}, |z| \geq 1, |\kappa| < 1$ ;

$$z - z^{-1} = \lambda(\kappa - \kappa^{-1}) \tag{6.3.2a}$$

$$z + z^{-1} = -\frac{1}{v_h}(\kappa + \kappa^{-1}), \quad v_h < 0. \tag{6.3.2b}$$

Assume that there is an eigenvalue $(z, \kappa)$ . Then $|z| \geq 1$ and $|\kappa| < 1$ implies

$$\operatorname{sign} \operatorname{Im} (\kappa + \kappa^{-1}) = -\operatorname{sign} \operatorname{Im} (\kappa - \kappa^{-1}) \tag{6.3.3a}$$

$$\operatorname{sign} \operatorname{Im} (z + z^{-1}) = \operatorname{sign} \operatorname{Im} (z - z^{-1}) . \tag{6.3.3b}$$

Comparing Eq.(6.3.3) with Eq.(6.3.2), one concludes that $\operatorname{Im} (\kappa - \kappa^{-1}) = 0$ . Similarly, by working with the real part, one finds that $\operatorname{Re} (\kappa - \kappa^{-1}) = 0$ . This shows that $\kappa = 1$ which contradicts the definition of eigenvalue. Thus we may conclude that there are no eigenvalues.

The absorption property of Eq.(6.3.1) is characterized by the reflection coefficient

$$r_{B_d} = \left| \frac{B_d(z, \kappa_-)}{B_d(z, \kappa_+)} \right| ,$$

where $|z| = |\kappa_\pm| = 1$ . Working with the dispersion relation Eq.(6.1.4) and the boundary condition Eq.(6.3.1), we find that

$$r_{B_d} = \prod_{h=1}^{d} \left| \frac{\lambda v_h + \sqrt{1 + (\lambda^2 - 1)/\cos^2 \tau \Delta t}}{\lambda v_h - \sqrt{1 + (\lambda^2 - 1)/\cos^2 \tau \Delta t}} \right| , \tag{6.3.4}$$

where $v_h \in [-1,0)$ and $\tau \in (-\tau_c, \tau_c)$ . From this it follows that

$$r_{B_d} < 1, \quad \text{and} \quad \lim_{d \to +\infty} r_{B_d} = 0 .$$

So far we have only studied the difference scheme without any reference to the differential equation. Next, we examine the solution of the leap-frog scheme with boundary condition Eq.(6.3.1) as $\Delta t \to 0$ (with $\lambda$ fixed), in order to see how well it approximates the solution of the differential equation. For this purpose, we consider the first-order boundary condition with $v = -1$ , tuned to the most rapid downstream speed. In this case, the reflection coefficient Eq.(6.3.4) becomes

$$\begin{aligned} r_1 &= \frac{(\lambda^2 - 1)\sin^2 \tau\Delta t}{(1 + \sqrt{1 + (\lambda^2 - 1)/\cos^2 \tau\Delta t})^2 \cos^2 \tau\Delta t} \\ &= O((\tau\Delta t)^2), \quad \text{as } \tau\Delta t \to 0 . \end{aligned}$$

This shows that as $\tau\Delta t \to 0$ ( $\lambda$ fixed), the amplitude of the reflected waves is of order $O((\tau\Delta t)^2)$ . In other words, the artificial boundary condition only introduces an error of order $O((\tau\Delta t)^2)$ in the solution of the leap-frog scheme. This error is acceptable because the truncation error of the leap-frog scheme itself is of order $O((\tau\Delta t)^2)$ . Therefore, in approximating the solution of the differential equation Eq.(6.0.1), no higher-order absorbing boundary condition is necessary.

We report some numerical results. In the computations, the leap-frog scheme is solved on the interval $[0,1]$ , with $\Delta x = \frac{1}{100}, \lambda = \frac{1}{2}$ . For an initial value, the wave packet

$$u_j^0 = e^{-100(\Delta x j - 0.5)^2} \cos \xi \Delta x j$$

is located at the middle of the interval $x = 0.5$ . On the right hand side boundary $x = 1$ , the condition $u_{100}^n = 0$ is imposed.

Fig.6.3.1 shows the results for the initial wave packet $\xi\Delta x = 0$ . The wave packet propagates at group velocity $C_x = -1$, travelling towards the boundary $x = 0$. When it reaches the boundary, different reflections occur when different boundary conditions are used in $x = 0$. In Fig.6.3.1(a), a perfectly reflecting condition $u_0^n = 0$ is applied, resulting in total reflection which has the sawtoothed shape travelling at group velocity $C_x = 1$. In Fig.6.3.1(b), a first-order absorbing condition is used with $v = -1$, which annihilates the incident wave packet; no reflection is detected. A first-order absorbing condition is also used in Fig.6.3.1(c), but this time, $v = -.5$. Reflection is not totally eliminated but the amplitude is significantly reduced as compared to Fig.6.3.1(a).

Fig.6.3.2 shows the reflection of the wave packet $\xi\Delta x = .35$. This wave packet travels at group velocity $C_x = -.507$. Total reflection (Fig.6.3.2(a)), partial reflection (Fig.6.3.2(b)) and nearly total absorption (Fig.6.3.2(c)) occur when the boundary conditions $u = 0$, the first-order condition Eq.(6.3.1) with $v = -1$ and the first-order condition Eq.(6.3.1) with $v = -.5$ are applied at $x = 0$, in that order.

### 6.3.2 THE KLEIN-GORDON EQUATION.

As another example, we consider the one-dimensional Klein-Gordon equation

$$u_{tt} - u_{xx} + u = 0 .$$

When it is discretized by a central differencing scheme, the result is

$$P(E_t, E_x)u_j^n = \left\{(E_t - 2 + E_t^{-1}) - \lambda^2(E_x - 2 + E_x^{-1}) + (\Delta t)^2\right\}u_j^n = 0 . \quad (6.3.5)$$

The von Neumann stability condition requires that $\left(\frac{\Delta t}{2}\right)^2 + \lambda^2 \le 1$. By substituting $u_j^n = e^{i(\tau\Delta tn + \xi\Delta xj)}$ into Eq.(6.3.5), the dispersion relation is found

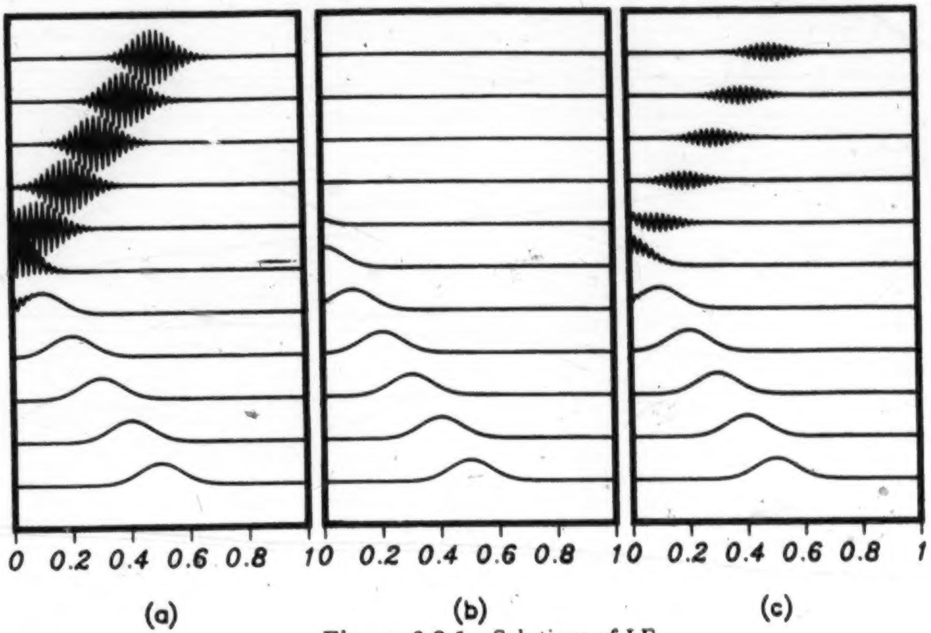

**Figure 6.3.1** Solutions of LF

Reflections at boundary $x = 0$ with different boundary conditions. Initial wave packet $\xi\Delta x = 0$ . (a) $u_0^n = 0$ . (b) Eq.(6.3.1) with $v = -1$ . (c) Eq.(6.3.1) with $v = -.5$ .

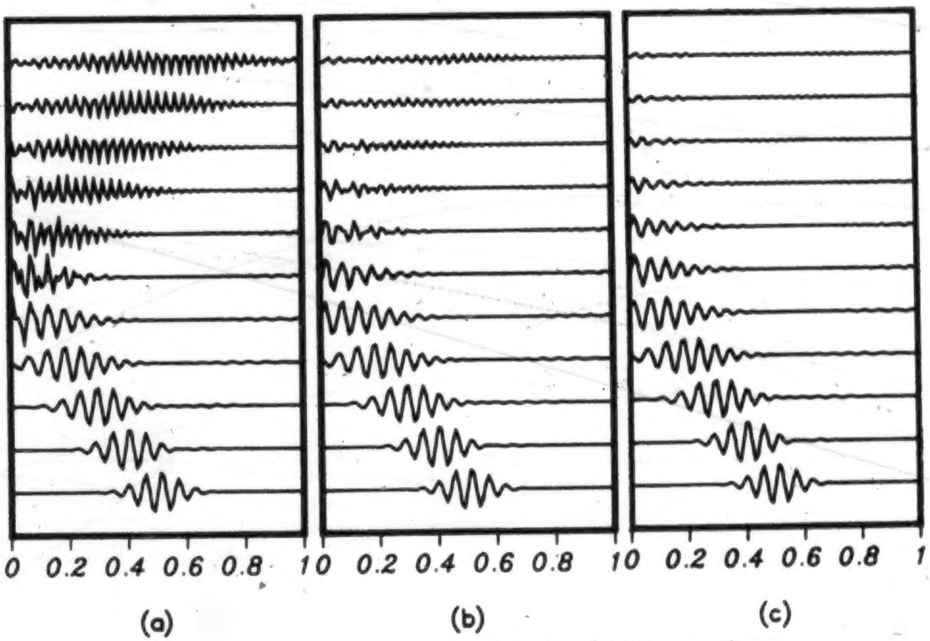

**Figure 6.3.2** Solutions of LF

Reflections at boundary $x = 0$ with different boundary conditions. Initial wave packet $\xi\Delta x = .35$ . (a) $u_0^n = 0$ . (b) Eq.(6.3.1) with $v = -1$ . (c) Eq.(6.3.1) with $v = -.5$ .

to be

$$\sin^2 \frac{\tau \Delta t}{2} = \left(\frac{\Delta t}{2}\right)^2 + \lambda^2 \sin^2 \frac{\xi \Delta x}{2} .$$

This is illustrated in Fig.6.3.3.

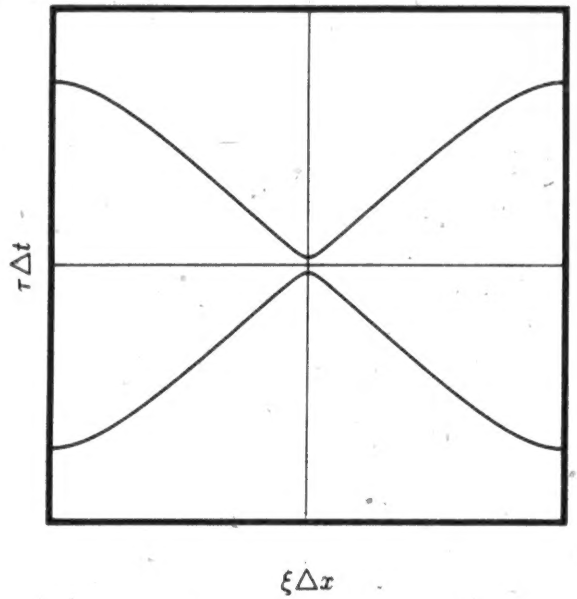

**Figure 6.3.3** Dispersion Relation for the Discrete Klein-Gordon Equation

According to Thm.6.2.1, the necessary form of a stable $\mathcal{L}^\infty_\varepsilon$ -best absorbing boundary condition of degree $d$ is given by

$$B_d(E_t, E_x)u_j^n = \prod_{h=1}^{d} \left\{ \lambda(E_x - E_x^{-1}) + v_h(E_t - E_t^{-1}) \right\} u_j^n = 0 , \qquad (6.3.6)$$

where $v_h \in (-1, -\varepsilon]$ (if $\left(\frac{\Delta t}{2}\right)^2 + \lambda^2 < 1$). In order to show that the condition Eq.(6.3.6) is stable, we consider the common solutions of

$$\begin{aligned} (z - 2 + z^{-1}) - \lambda(\kappa - 2 + \kappa^{-1}) + (\Delta t)^2 &= 0 \\ v_h(z - z^{-1}) + \lambda(\kappa - \kappa^{-1}) &= 0 . \end{aligned} \qquad (6.3.7)$$

If Eq.(6.3.7) admits a solution $(z,\kappa) \in \mathbb{C} \times \mathbb{C}, |z| \geq 1, |\kappa| < 1$, then the same procedure as in Example 6.3.1 leads to $\operatorname{Im}(\kappa - \kappa^{-1}) = 0$, which in turn implies $z, \kappa \in \mathbb{R}$. If $z \geq 1$, then $\kappa \leq 0$ from the second equation in Eq.(6.3.7). This shows that $z - 2 + z^{-1} \geq 0$ and $\kappa - 2 + \kappa^{-1} \leq 0$, contradicting the first equation in Eq.(6.3.7). On the other hand, if $z \leq -1$, then $\kappa \geq 0$. Therefore $z - 2 + z^{-1} \leq -4$ and $\lambda(\kappa - 2 + \kappa^{-1}) \geq 0$. Again this contradicts the first equation in Eq.(6.3.7) because $(\triangle t)^2 - 4 < 0$, by the von Neumann condition. Thus we have shown that there are no eigenvalues.

Each factor in Eq.(6.3.6) is equivalent to

$$\left\{ \frac{E_x - E_x^{-1}}{\triangle x} + v_h \frac{E_t - E_t^{-1}}{\triangle t} \right\} u_j^n = 0 .$$

This can be regarded as the central difference approximation of

$$u_x + v_h u_t = 0$$

which is the first-order absorbing boundary condition of Eq.(5.1.1). This shows that the absorbing boundary conditions derived from both the differential equation and the difference equation are consistent.

We now present some numerical results. The Klein-Gordon equation is discretized by the central differencing scheme Eq.(6.3.5) on $[0,1]$ with $\triangle x = \frac{1}{100}, \lambda = .5$. The initial function is given by a Dirac delta-like pulse concentrated at $x = .5$, i.e.,

$$u_j^0 = 0, \quad j \neq 50, \qquad u_{50}^0 = 15 .$$

Since the discrete Fourier transform of the above $u^0$ is a constant [ViBo82], the initial function is composed of wave modes of all wave numbers $\xi \triangle x \in [-\pi, \pi]$ (with the same amplitude). After a short time, the initial pulse will disperse into waves of different wave lengths, travelling at different group velocities. We are concerned with the reflections when these waves strike the boundary $x = 0$. The reflection is obtained by comparing the solution in $[0, 1]$ with the free-space solution which is computed by solving Eq.(6.3.5) in a much larger domain $[-4, 1]$. Fig.6.3.4 shows the total reflection with boundary condition $u_0^n = 0$ at $x = 0$, at time $t = 1.5$. Note the wave front has arrived at $x \approx 1$, because it propagates with group velocity $C_x \approx -1$ for $t \in [0, .5]$ before it hits the boundary $x = 0$, and at the speed $C_x \approx 1$ for $t \in [.5, 1.5]$ after it is reflected. Thus at time $t = 1.5$, the wave front has travelled a total distance of approximately 1.5, arriving at $x \approx 1$.

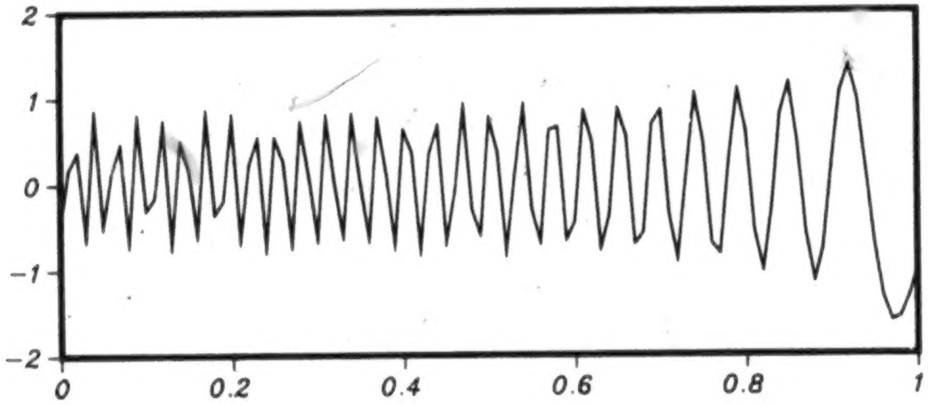

**Figure 6.3.4** Total reflections
Reflections at $t = 1.5$

Fig.6.3.5 shows the absorption of the first-order boundary condition Eq.(6.3.6) with various parameters $v$. Note that the part of the wave travelling at speed $C_x = v$ is annihilated exactly.

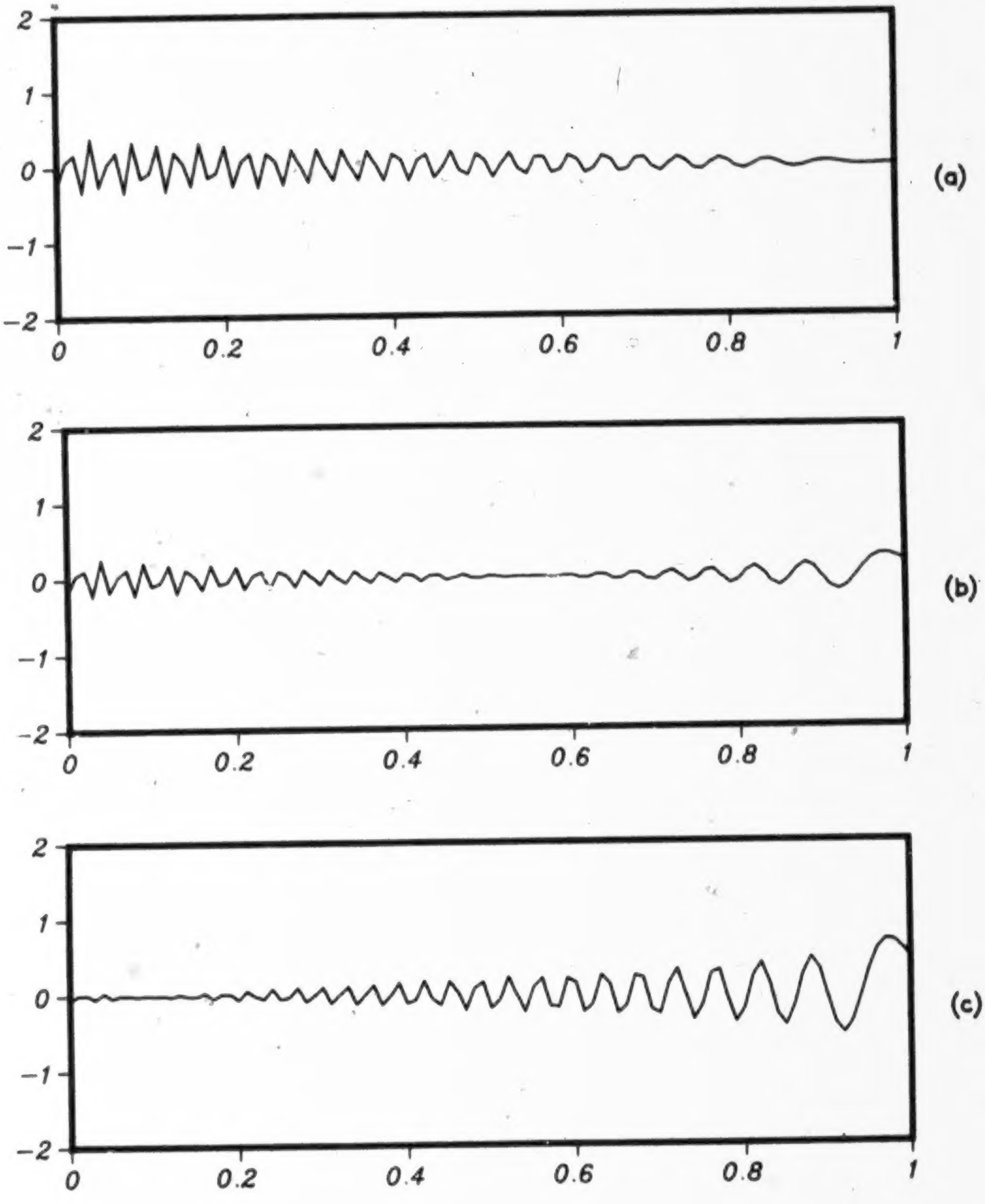

**Figure 6.3.5** Reflections of first-order boundary conditions
(a) $v = 1$ , (b) $v = .7$ , (c) $v = .4$ .

Fig.6.3.6 and Fig.6.3.7 show the reflections from the second-order and the third-order boundary conditions, with different parameters.

Next, we illustrate how the parameters can be made time dependent to improve the absorption properties of a boundary condition. In previous computations, the intial value is concentrated at $x = .5$ . Hence each wave mode has to travel an exact distance of .5 to reach the boundary $x = 0$ , independent of the frequencies. At a given time $t_0$ , the wave that hits the boundary must propagate at the exact speed $C_x = .5/t_0$ , which can be annihilated by tuning $v = C_x = .5/t_0$ . This suggests that the reflection can be totally eliminated if $v = \frac{.5}{t}$ in the first-order boundary condition. Thus we consider

$$\lambda(u_{j+1}^n - u_{j-1}^n) + v_n(u_j^{n+1} - u_j^{n-1}) = 0\,, \tag{6.3.8}$$

where $v_n = \frac{.5}{n\Delta t}$ .

Fig.6.3.8 shows that this is indeed true. This result should be compared with Fig.6.3.5 where the parameter $v$ is kept constant. The success of this technique lies, of course, in *a priori* information that the waves are all initiated from a fixed distance from the boundary. Thus the variable parameter technique can be successful for problems where the disturbance is caused by some forcing term which is fixed in space.

Finally, we illustrate weak instabilities of absorbing boundary conditions. As shown before, Eq.(6.3.6) admits a generalized eigenvalue $(z, \kappa) = (1, 1)$ , which may cause a weak instability. In our experiments, this generalized eigenvalue has no observed effect on the first- and the second-order boundary conditions. This agrees with the observation made by Higdon [Higd86a, 87]. For the third-order boundary conditions, the weak instability is only observed when initial wave packets have large wave length ( $\xi\Delta x \approx 0$ ). The effect of the generalized eigenvalue

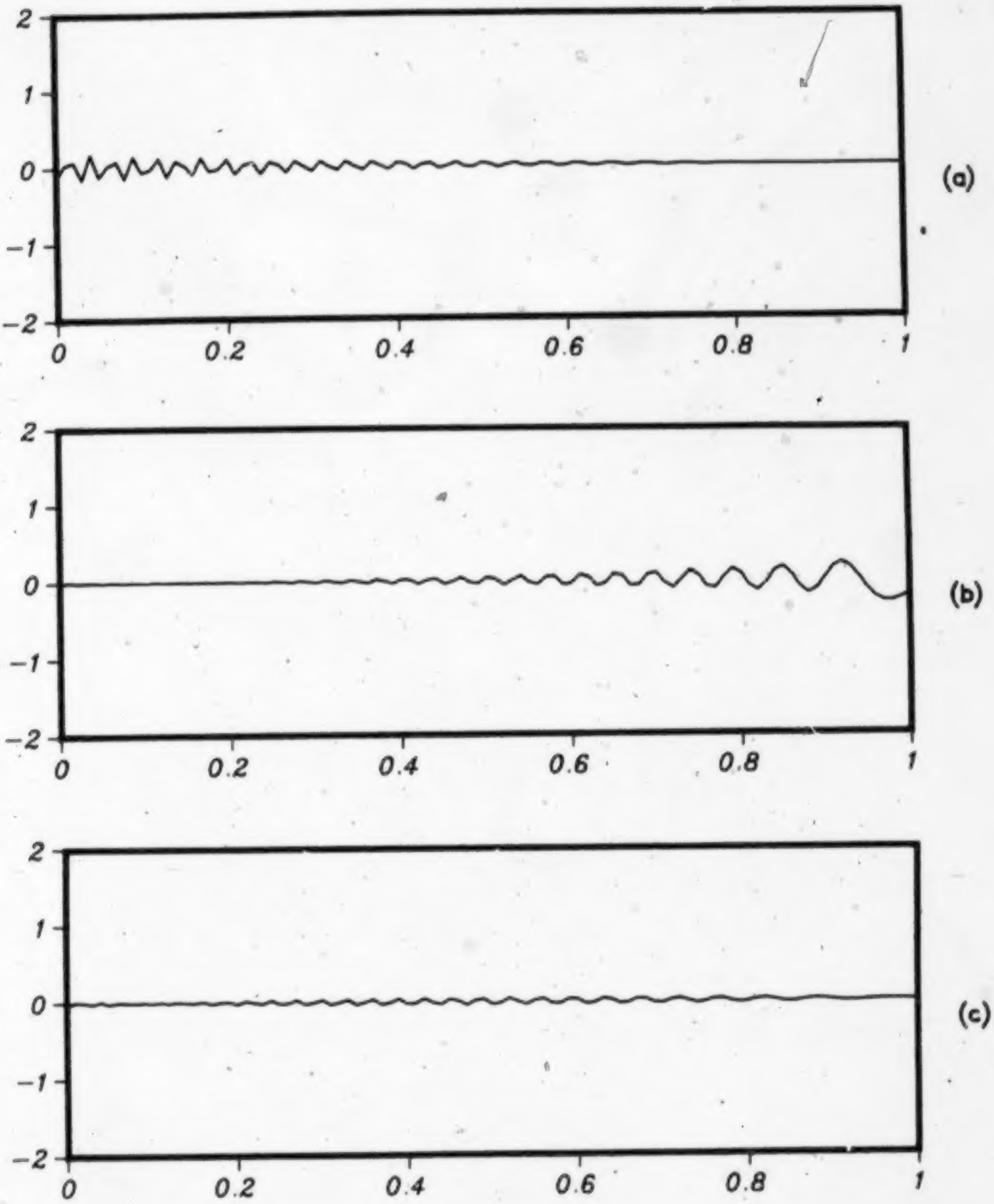

**Figure 6.3.6** Reflections of second-order boundary conditions (a) $v_1 = v_2 = 1$ , (b) $v_1 = v_2 = .4$ , (c) $v_1 = .4,\ v_2 = 1$ .

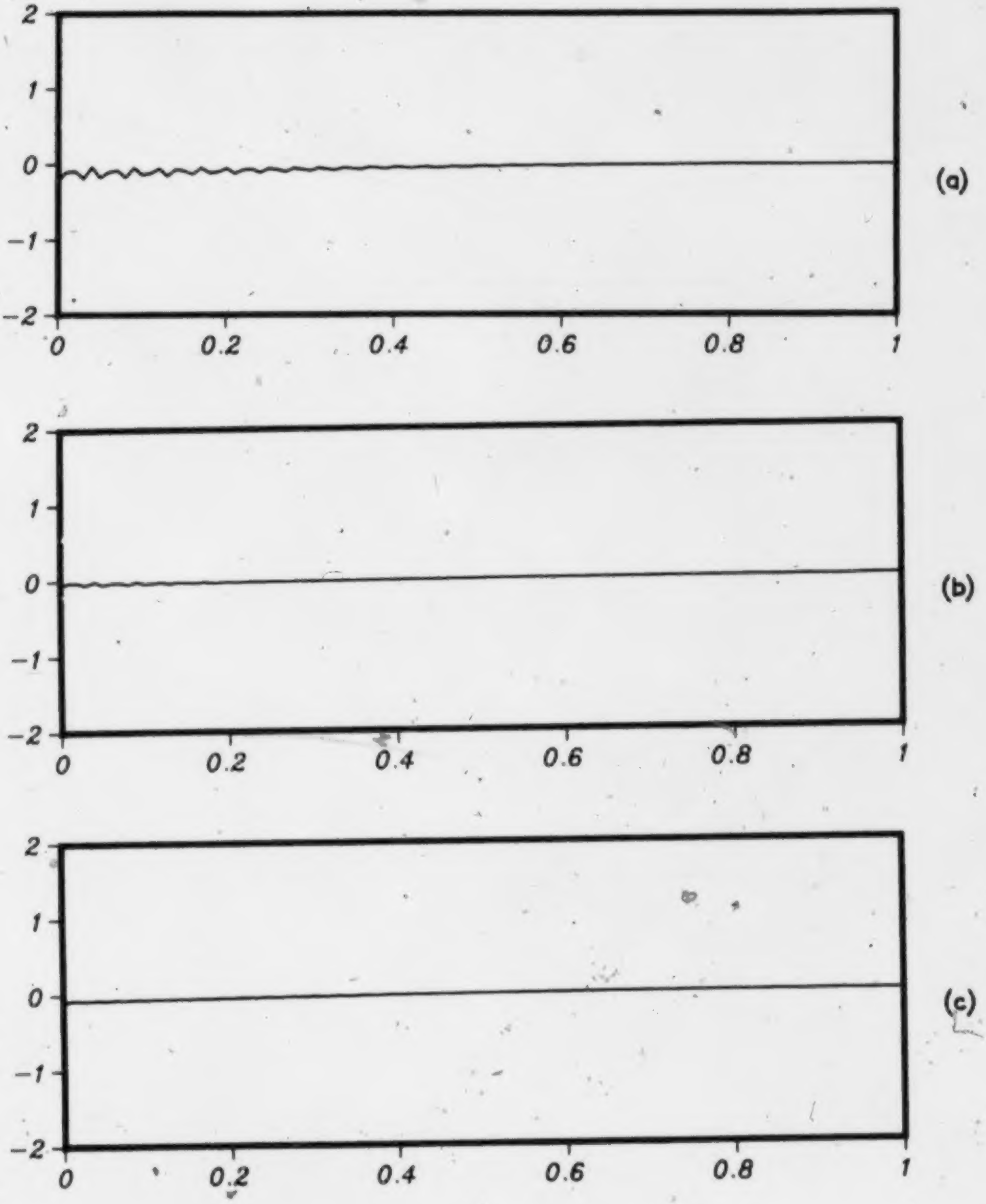

**Figure 6.3.7** Reflections of third-order boundary conditions
(a) $v_1 = v_2 = v_3 = 1$ , (b) $v_1 = v_2 = v_3 = .7$ , (c) $v_1 = .4,\ v_2 = .7,\ v_3 = 1$ .

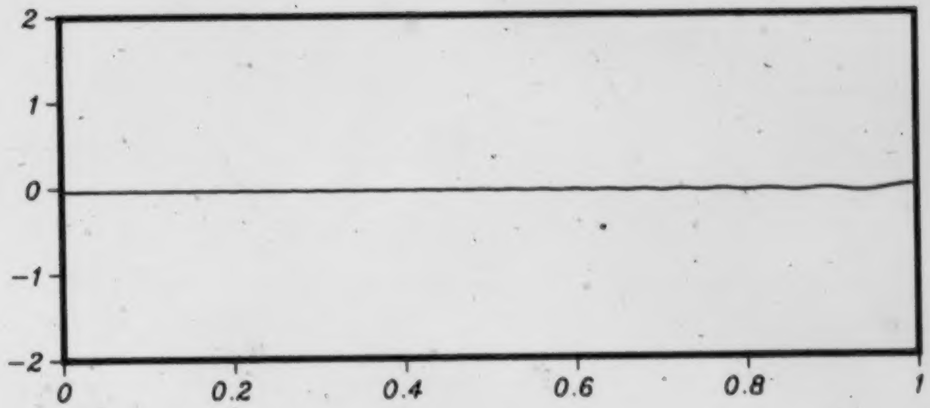

**Figure 6.3.8** Total absorption
In the first-order condition, the parameter $v$ is a time dependent variable.

is negligible for waves with large wave numbers. The reason for this is that the troublesome mode is $(\tau\triangle t, \xi\triangle x) = (0,0)$ which is the only generalized eigenvalue.

In Fig.6.3.9, a third-order boundary condition is used. In Fig.6.3.9(a) and Fig.6.3.9(b) the parameters are $v_1 = v_2 = v_3 = 0$, which yields a reflection coefficient of amplitude 1, hence is perfectly reflecting. Fig.6.3.9(a) shows the reflection of an initial wave packet

$$u^0(x) = e^{-100(x-.5)^2} \cos \xi x, \tag{6.3.9}$$

with $\xi\triangle x = .05$. The instability is visualizable from the nonzero amplitude of reflection after the main disturbance has long passed (the exact reflection should be 0 at $t = 2$ ). In Fig.6.3.9(b), initial disturbance is given by Eq.(6.3.9) with $\xi\triangle x = .2$. No instability is observed in this case.

The instability is even weaker when the boundary condition is highly absorbing. Fig.6.3.9(c) confirms this point. In this case, $\xi\Delta x = .05$ , the parameters are $v_1 = v_2 = v_3 = -1$ . The corresponding boundary condition has the same generalized eigenvalue but its reflection coefficient is very small.

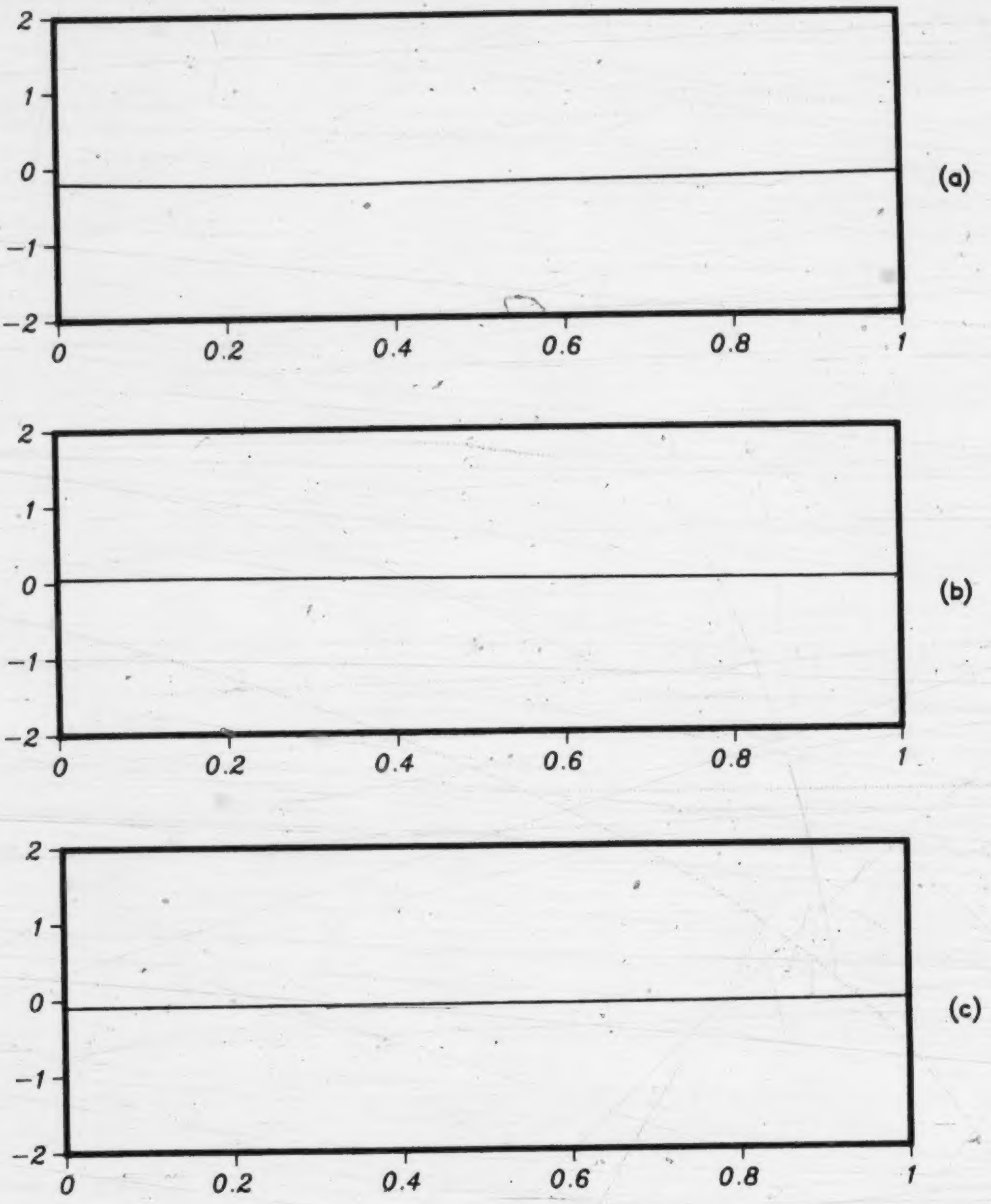

**Figure 6.3.9** Effects of generalized eigenvalues
t=2; (a) $v_1 = v_2 = v_3 = 0$ , and $\xi\Delta x = .05$ (b) $v_1 = v_2 = v_3 = 0$ , and $\xi\Delta x = .2$ , (c) $v_1 = v_2 = v_3 = -1$ , and $\xi\Delta x = .05$ .

# PART II POLYNOMIAL PRECONDITIONING TECHNIQUES

# CHAPTER 7

# POLYNOMIAL PRECONDITIONING FOR CONJUGATE GRADIENT METHODS

This chapter is a survey of the basic theory of polynomial preconditioning of conjugate gradient methods. In section one, we review the basic preconditioning conjugate gradient algorithm and discuss its convergence behavior. Section two deals with polynomial preconditioning techniques with various selections of polynomial operators.

## 7.1 Preconditioned Conjugate Gradient Algorithm

Consider the linear system

$$Au = b\,, \tag{7.1.1}$$

where $A$ is an $n \times n$ symmetric and positive definite matrix. Let $M$ be an arbitrary $n \times n$ symmetric and positive definite matrix. The preconditioned conjugate gradient (PCG) algorithm for the iterative solutions of Eq.(7.1.1) is given as follows [CGO76]:

PCG ALGORITHM.

For a given initial approximation $u_0$, find the initial residual by $r_0 = b - Au_0$, and the initial correction $d_0$ by $Md_0 = r_0$. Let the initial direction be $p_0 = d_0$.

For $j = 0, 1, 2, \dots$, compute Step 1 through Step 6 until some stopping criterion is satisfied:

Step 1. $\alpha_j = \langle r_j, d_j\rangle / \langle p_j, Ap_j\rangle$ ,

Step 2. $u_{j+1} = u_j + \alpha_j p_j$ ,

Step 3. $r_{j+1} = r_j - \alpha_j Ap_j$ ,

Step 4. Solve $Md_{j+1} = r_{j+1}$ for $d_{j+1}$ ,

Step 5. $\beta_j = \langle r_{j+1}, d_{j+1}\rangle / \langle r_j, d_j\rangle$ ,

Step 6. $p_{j+1} = d_{j+1} + \beta_j p_j$ .

In the above and hereafter, $\langle \cdot, \cdot \rangle$ stands for the Euclidean inner product. Setting $M = I$ (the identity matrix) results in the standard conjugate gradient algorithm of Hestenes and Stiefel [HeSt52].

It is easy to show from the PCG algorithm that the $j+1$ st iteration $u_{j+1}$ satisfies $u_{j+1} = u_j + u$ , for some $u \in K^{j+1}(A, M, r_0)$ , where $K^{j+1}(A, M, r_0)$ is the Krylov space defined by

$$K^{j+1}(A, M, r_0) = \operatorname{span}\{d_0, M^{-1}Ad_0, \dots, (M^{-1}A)^j d_0\}. \tag{7.1.2}$$

By induction, we find that the $j+1$ st iteration satisfies

$$u_{j+1} \in u_0 + K^{j+1}(A, M, r_0).$$

We denote by $u^*$ the exact solution of Eq.(7.1.1) and by $e_j = u_j - u^*$ the error at the $j$ th iteration. Then the preconditioned conjugate gradient algorithm is characterized by the minimization property, namely, that $u_j \in u_0 + K^j(A, M, r_0)$ is determined to minimize the $A$ -norm of the error $\|e_j\|_A = \langle Ae_j, e_j\rangle^{1/2}$ .

7.1.1. THEOREM [CGO76]. *If $u_j$ is computed from the PCG algorithm, then*

$$\|e_j\|_A = \min_{u \in u_0 + K^j} \|u - u^*\|_A .$$

**Proof:** See [CGO76]. □

Since $u_j \in u_0 + K^j(A, M, r_0)$, it follows that $u_j - u_0 \in K^j(A, M, r_0)$. Therefore,

$$\begin{aligned} u_j - u_0 &= \sum_{k=0}^{j-1} c_k (M^{-1}A)^k d_0 \\ &= \sum_{k=0}^{j-1} c_k (M^{-1}A)^{k+1} e_0 , \end{aligned}$$

where $c_k$ 's are some constants. This relation leads to $e_j = u_j - u^* = u_0 - u^* + u_j - u_0 = e_0 + \sum_{k=0}^{j-1} c_k (M^{-1}A)^{k+1} e_0$. In other words, there is some polynomial $p_j \in \Pi_j^1$, the class of polynomials of degree at most $j$ with $p_j(0) = 1$, such that

$$e_j = p_j(M^{-1}A)e_0, \qquad \text{or} \qquad r_j = p_j(M^{-1}A)r_0 , \tag{7.1.3}$$

where $r_j$ is the residual at the $j$ th iteration.

Let $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ be the eigenvalues of the preconditioned matrix $M^{-1}A$ and $\phi_1, \phi_2, \ldots, \phi_n$ be the corresponding $A$ -orthonormal eigenvectors. When $e_0$ is expressed in terms of $\phi_k$ 's,

$$e_0 = \sum_{k=1}^{n} \gamma_k \phi_k , \tag{7.1.4}$$

and inserted into Eq.(7.1.3), the result is

$$e_j = p_j(M^{-1}A)e_0 = \sum_{k=1}^{n} \gamma_k p_j(M^{-1}A)\phi_k = \sum_{k=1}^{n} \gamma_k p_j(\lambda_k)\phi_k .$$

Therefore the $A$ -norm of the error is

$$\|e_j\|_A = \langle Ae_j, e_j \rangle^{1/2} = \sqrt{\sum_{k=1}^{n} \gamma_k^2 p_j^2(\lambda_k)} .$$

7.1.2. COROLLARY. *If the iteration* $u_j$ *is computed from PCG, then the error* $e_j$ *satisfies*

$$\|e_j\|_A = \min_{p_j \in \Pi_j^1} \sqrt{\sum_{k=1}^{n} \gamma_k^2 p_j^2(\lambda_k)}\,,$$

*where* $\gamma_k$ *is the component of* $e_0$ *into the direction of* $\phi_k$ *, as given in Eq.(7.1.4).*

PROOF: Follows from Thm.7.1.1. ☐

The above corollary shows that the polynomial in Eq.(7.1.3) produced by PCG is the best least squares approximation to $0$ on the discrete set $\{\lambda_1, \cdots, \lambda_n\}$ with weights $\{\gamma_1^2, \cdots, \gamma_n^2\}$ . Therefore, the eigenvalue distribution of the preconditioned system $M^{-1}A$ plays an important role in the performance of the PCG algorithm.

The error reduction of the PCG algorithm can be estimated by the formula given in Cor.7.1.2. Let us denote the spectrum of the matrix $M^{-1}A$ by $\sigma(M^{-1}A) = \{\lambda_1, \cdots, \lambda_n\}$ . It follows from Cor.7.1.2 that

$$\begin{aligned}
\|e_j\|_A &= \min_{p_j \in \Pi_j^1} \sqrt{\sum_{k=1}^{n} \gamma_k^2 p_j^2(\lambda_k)} \\
&\le \min_{p_j \in \Pi_j^1} \max_{\lambda \in \sigma(M^{-1}A)} |p_j(\lambda)| \sqrt{\sum_{k=1}^{n} \gamma_k^2} \\
&\le \min_{p_j \in \Pi_j^1} \max_{\lambda \in \sigma(M^{-1}A)} |p_j(\lambda)|\, \|e_0\|_A\,.
\end{aligned}$$

This leads to the following lemma.

7.1.3. LEMMA. *Let* $u_j$ *be generated by PCG. Then*

$$\frac{\|e_j\|_A}{\|e_0\|_A} \le \min_{p_j \in \Pi_j^1} \max_{\lambda \in \sigma(M^{-1}A)} |p_j(\lambda)|\,. \tag{7.1.5}$$

An immediate consequence of Lem.7.1.3 is

7.1.4. COROLLARY. *If the number of nonzero elements in the set* $\{\gamma_1, \cdots, \gamma_n\}$ *is* $k$ *, then the exact solution of Eq.(7.1.1) is obtained after at most* $k$ *iterations, i.e.,* $e_k = 0$ .

PROOF: Without loss of generality, let $\gamma_1, \cdots, \gamma_k \neq 0$ , and $\gamma_{k+1}, \cdots, \gamma_n = 0$ . We define a polynomial $p_k^0 \in \Pi_k^1$ by

$$p_k^0 = \prod_{j=1}^{k} \left(1 - \frac{\lambda}{\lambda_j}\right),$$

where $\lambda_j \in \sigma(M^{-1}A)$ . Since $\gamma_{k+1}, \ldots, \gamma_n = 0$ , Eq.(7.1.5) becomes

$$\|e_k\|_A \leq \|e_0\|_A \min_{p_j \in \Pi_j^1} \max_{\lambda \in \{\lambda_1, \ldots, \lambda_k\}} |p_j(\lambda)| .$$

The right hand side of the above is bounded by $\|e_0\|_A \max_{\lambda \in \{\lambda_1, \ldots, \lambda_k\}} |p_k^0(\lambda)| = 0$ , by definition of $p_k^0$ . □

The above result holds only in exact arithmetic. In the presence of roundoff error, the PCG algorithm generates a polynomial that is slightly different from the best least squares polynomial. In this case, the exact solution cannot be obtained after a finite number of iterations, but rather the error is reduced at each iteration. Another conclusion from the above corollary is that if the eigenvalues of $M^{-1}A$ cluster around a few points of $\sigma(M^{-1}A)$ , say $\lambda_1, \ldots, \lambda_k$ , then the error $e_j$ will be very small after $k$ iterations, because as in the proof of Cor.7.4.4, $|p_j^0(\lambda)|$ is very small for all $\lambda \in \{\lambda_1, \ldots, \lambda_k\}$ .

It is possible to find an explicit estimate of the error reduction in the PCG iteration; the following theorem can be found in [JMP83].

7.1.5. THEOREM [JMP83]. *Let* $u_j$ *be computed from PCG, and let* $\kappa = \kappa(M^{-1}A) = \lambda_n/\lambda_1$. $\lambda_1, \lambda_n$ *are the smallest and the largest eigenvalues of* $M^{-1}A$ , *respectively. Then*

$$\frac{\|e_j\|_A}{\|e_0\|_A} \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^j .$$

PROOF: When $\sigma(M^{-1}A)$ is replaced by $[\lambda_1, \lambda_n]$ in Eq.(7.1.5), the result is

$$\frac{\|e_j\|_A}{\|e_0\|_A} \le \min_{p_j \in \Pi_j^1} \max_{\lambda \in [\lambda_1, \lambda_n]} |p_j(\lambda)| .$$

Let $T_j$ be the Chebyshev polynomial of degree $j$ . Then the minimum of the right hand side is achieved by

$$p_j(\lambda) = T_j(\frac{2\lambda - \lambda_n - \lambda_1}{\lambda_n - \lambda_1})\, T_j^{-1}(\frac{-\lambda_n - \lambda_1}{\lambda_n - \lambda_1}) .$$

Using the properties of the Chebyshev polynomial, we arrive at

$$\begin{aligned} \frac{\|e_j\|_A}{\|e_0\|_A} &\le |T_j^{-1}(\frac{-\lambda_n - \lambda_1}{\lambda_n - \lambda_1})| \\ &\le 2\frac{\nu^j}{1+\nu^{2j}} \\ &\le 2\nu^j , \end{aligned}$$

where

$$\nu = \frac{\sqrt{\frac{\lambda_n}{\lambda_1}} - 1}{\sqrt{\frac{\lambda_n}{\lambda_1}} + 1} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} ,$$

thus completing the proof. □

The bound in Thm.7.1.5 is often quite pessimistic, because in the proof of the theorem, we have replaced the discrete set $\sigma(M^{-1}A)$ by the interval $[\lambda_1, \lambda_n]$ .

7.1.6. DEFINITION. *The condition number of the preconditioned system* $M^{-1}A$ *is defined by* $\kappa = \kappa(M^{-1}A) = \lambda_n/\lambda_1$ . *Its convergence factor is defined by*

$$\nu = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \qquad \kappa = \kappa(M^{-1}A) .$$

Thm.7.1.5 indicates that

$$\|e_j\|_A \le 2\nu^j \|e_0\|_A ,$$

that is, PCG converges rapidly if the convergence factor is small or, equivalently, if the condition number is small.

## 7.2 Polynomial Preconditionings

It is clear from the previous section that the performance of the PCG method largely depends on the choice of the preconditioning operator $M$. The important considerations in the selection of $M$ are that $M^{-1}$ should be a good approximation to $A^{-1}$ in some sense, so that $\kappa(M^{-1}A)$ is small, and that the preconditioning step 4 of the PCG algorithm can be conveniently carried out, i.e., $d = M^{-1}r$ should be easily computed. The latter requirement is satisfied if the preconditioning operator is chosen so that $M^{-1}$ is a polynomial of $A$. In this section, we consider various choices of the polynomials and analyze their convergence behaviors.

### Truncated Neumann Series

A simple choice of $M$ is based on a Neumann series for approximating a matrix inverse [DGR79]. In this technique, $A$ is symmetrically scaled to have unit diagonal elements, hence it can be expressed as

$$A = I - G.$$

If the spectral radius $\rho(G)$ of $G = I - A$ is less than $1$, then $A^{-1}$ can be computed by the infinite series

$$A^{-1} = \sum_{j=0}^{\infty} G^j.$$

An approximation to $A^{-1}$ is simply a truncated form of this series; that is,

$$M^{-1} = \sum_{j=0}^{m} G^j. \tag{7.2.1}$$

With this choice of $M$ , the preconditioning step 4, $d_{j+1} = M^{-1}r_{j+1}$ , can be easily carried out by Horner's rule; no powers of $G$ need to be formed explicitly.

In some cases, the spectral radius of $G = I - A$ may be greater than $1$ , so that the above form is not valid because $M$ defined through Eq.(7.2.1) may not be positive definite as required by the PCG algorithm. However, Eq.(7.2.1) can be modified to apply to such cases. Thus we define

$$G = I - \frac{1}{\omega}A, \qquad \omega > \frac{1}{2}\|A\|_2 . \tag{7.2.2}$$

It can be verified that $G$ defined in Eq.(7.2.2) satisfies $\rho(G) < 1$ . In the following, $G$ is defined by Eq.(7.2.2) whenever the spectral radius of $I - A$ is greater than $1$ .

To insure that the PCG algorithm converges, we have to verify that the preconditioning operator $M$ of Eq.(7.2.1) is symmetric and positive definite.

7.2.1. THEOREM. *If $A$ is symmetric and positive definite, then $M^{-1}$ of Eq.(7.2.1) is also symmetric and positive definite.*

PROOF: This is a modification of the proof given in [DGR79].

Since $A$ is symmetric by assumption, so are $G$ and its powers. The symmetry of $M^{-1}$ follows immediately from its definition.

By a direct calculation, $I - \frac{1}{\omega}M^{-1}A = G^{m+1}$ , where $I$ is the identity. Thus $\rho(G) < 1$ implies that the eigenvalues of $\frac{1}{\omega}M^{-1}A$ are positive. This in turn shows that the eigenvalues of $M^{-1}$ are all positive because $A$ is symmetric and positive definite. This shows that $M^{-1}$ is positive definite. ☐

The next theorem gives the condition number of the Neumann series preconditioning.

7.2.2. **Theorem** [Adam85]. *Let the eigenvalues of* $G$ *be ordered as*

$$-1 < \tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_n < 1 .$$

*Then the condition number of the preconditioned system by the Neumann series is*

$$\kappa(M^{-1}A) = \begin{cases} \dfrac{1-\tilde{\lambda}_1^{m+1}}{1-\tilde{\lambda}_n^{m+1}}, & m+1 \quad \text{odd} \\[2ex] \dfrac{1-\tilde{\mu}^{m+1}}{1-\tilde{\lambda}_n^{m+1}}, & m+1 \quad \text{even, and} \quad |\tilde{\lambda}_n| \geq |\tilde{\lambda}_1| \\[2ex] \dfrac{1-\tilde{\mu}^{m+1}}{1-\tilde{\lambda}_1^{m+1}}; & m+1 \quad \text{even, and} \quad |\tilde{\lambda}_n| \leq |\tilde{\lambda}_1| , \end{cases}$$

*where* $\mu = \min\{|\tilde{\lambda}_1|, \cdots, |\tilde{\lambda}_n|\}$ .

**Proof:** From the proof of Thm.7.2.1, $M^{-1}A = \omega(I - G^{m+1})$ . Therefore $\kappa(M^{-1}A) = \kappa(I - G^{m+1})$ . The rest of the proof follows easily by examining the extreme points of the function $1 - \lambda^{m+1}$ . □

Note that even though $\lim_{m\to\infty} \kappa(M^{-1}A) = 1$ , $\kappa(M^{-1}A)$ is not a monotonic function of $m$ . Increasing $m$ may in fact increase the condition number, hence slowing down the convergence. This difficulty can be overcome by the following methods.

### Minimax Polynomials

To accelerate the convergence, we may introduce some parameters $c_j$ into the Neumann truncated series to obtain a polynomial

$$M^{-1} = \sum_{j=0}^{m} c_j G^j .$$

Since $G = I - \frac{1}{\omega}A$ , the above is equivalent to a polynomial in $A$ , i.e.,

$$M^{-1} = p_m(A) , \tag{7.2.3}$$

for some polynomial $p_m \in \Pi_m$. The polynomial $p_m$ can be chosen so that the eigenvalue distribution of the preconditioned system $M^{-1}A$ is more favorable for the CG method. One criterion is to reduce the condition number of $M^{-1}A$. Ideally, we should select $p_m$ such that the expression

$$\kappa(M^{-1}A) = \frac{\max\limits_{\lambda \in \sigma(A)} \lambda p_m(\lambda)}{\min\limits_{\lambda \in \sigma(A)} \lambda p_m(\lambda)} \tag{7.2.4}$$

is minimized over all polynomials of degree at most $m$. This process, however, requires the knowledge of the spectrum $\sigma(A)$, which is generally unknown. We therefore replace $\sigma(A)$ by $[\lambda_1, \lambda_n]$ and turn to the problem of finding the minimum value $p_m \in \Pi_m$ of the expression

$$\frac{\max\limits_{\lambda \in [\lambda_1, \lambda_n]} \lambda p_m(\lambda)}{\min\limits_{\lambda \in [\lambda_1, \lambda_n]} \lambda p_m(\lambda)},$$

which is an upper bound of Eq.(7.2.4). The above expression is referred to as the *condition number* of the polynomial $\lambda p_m(\lambda)$ and is denoted by $\kappa(\lambda p_m)$.

7.2.3. THEOREM [JMP83]. *Let $q_m \in \Pi_m$ be defined by*

$$q_m(\lambda) = \frac{1}{\lambda}\left(1 - \frac{T_{m+1}\left(\frac{2\lambda - \lambda_n - \lambda_1}{\lambda_n - \lambda_1}\right)}{T_{m+1}\left(\frac{-\lambda_n - \lambda_1}{\lambda_n - \lambda_1}\right)}\right). \tag{7.2.5}$$

*Then its condition number satisfies*

$$\kappa(\lambda q_m) \leq \kappa(\lambda p_m) \quad \text{for all} \quad p_m \in \Pi_m .$$

PROOF: See [JMP83]. □

The above theorem indicates that by choosing $q_m$ as defined above, the upper bound on the condition number of the preconditioned system is minimized. Because of this, the polynomial $q_m$ is termed the *minimax polynomial.* The condition number decreases as the degree of the preconditioning polynomial increases.

By using a polynomial of high degree, the convergence of the PCG can be accelerated. The following theorem describes exactly how much the convergence factor can be improved.

7.2.4. THEOREM [JMP83]. *Let* $\nu$ *be the convergence factor of the CG with no preconditioning, i.e.,* $\nu = \frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}$ *, and let* $\nu_m$ *be the bound of the convergence factor of the PCG preconditioned by the polynomial* $q_m$ *of Eq.(7.2.5), i.e.,*

$$\nu_m = \frac{\sqrt{\kappa(\lambda q_m)} - 1}{\sqrt{\kappa(\lambda q_m)} + 1} .$$

*Then*

$$\nu_m = \nu^{m+1} . \tag{7.2.6}$$

PROOF: See [JMP83]. □

This theorem shows that one step of the PCG with the minimax polynomial of degree $m$ is as good as $m+1$ steps of the CG with no preconditioning. However, the computational work for the former is less because it requires fewer computations of inner products in steps 1 and 5 of the PCG algorithm.

### Least Squares Polynomials

The minimax polynomial discussed previously yields a minimum condition number for the preconditioned system. But a small condition number does not always mean fast convergence. The reason is that the condition number does not matter as much as the overall distribution of the eigenvalues. In this sense, the minimax polynomial is far from being the best polynomial, i.e., the one resulting in a preconditioned matrix that minimizes the total iteration count required by PCG. One obvious disadvantage of the minimax polynomial is that the extreme eigenvalues $\lambda_1$, $\lambda_n$ must be approximated very accurately, because the coefficients of the polynomial are very sensitive to $\lambda_1, \lambda_n$ ; poor approximations of the latter may significantly deteriorate the convergence.

An alternative is to use the best least squares polynomials as described in [JMP83] and [Saad85]. The advantage of this approach is that the two extreme eigenvalues need not be approximated very accurately. If $a$, $b$ are approximations of $\lambda_1, \lambda_n$ respectively, then $a$ needs not even be positive. In the least squares method, the preconditioning polynomial is the minimum value $p_m \in \Pi_m$ of the integral

$$\int_{\lambda_1}^{\lambda_n} w(\lambda)(1-\lambda p_m(\lambda))^2 d\lambda\,, \tag{7.2.7}$$

where $w$ is some weight function. Detailed discussions of the effects of different weight functions, and the computation of the coefficients of the best least squares polynomial can be found in [Saad85]. Here we will only be concerned with the convergence rate analysis for the case $w = 1$ .

7.2.5. THEOREM [JMP83]. *Let* $\nu = \dfrac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}$ *be the convergence factor of non-preconditioned CG, and* $\tilde{\nu}_m$ *be that of PCG preconditioned by* $M^{-1} = \tilde{q}_m(A)$ *where* $\tilde{q}_m$ *is the best least squares approximation of Eq.(7.2.7). Then*

$$\tilde{\nu}_m \le \sqrt{2}(m+2)\nu^{m+1}\,.$$

PROOF: See [JMP83] ☐

Note that the estimate of $\tilde{\nu}_m$ compares quite favorably with that given in Thm.7.2.4 for the minimax polynomial Numerical experiments have been conducted by Johnson *et al.*[JMP83] and Saad [Saad85]. Their results show that the least squares preconditioning may lead to faster convergence than the minimax preconditioning despite the fact that the estimate in Thm.7.2.5 is worse than that in Thm.7.2.4.

# CHAPTER 8

# A POLYNOMIAL PRECONDITIONING METHOD APPLIED TO THE SYSTEM $A^2u = b$

In this chapter, we propose a new polynomial preconditioning algorithm to solve linear systems of the form

$$A^2u = b \tag{8.0.1}$$

where $A$ is symmetric and positive definite. We will show that the new algorithm converges at a rate that is comparable to PCG applied to system $Au = b$, which is much better conditioned than Eq.(8.0.1). The technique used here belongs to a more general class known as *approximate polynomial preconditioning* ([Wong88], [WoJi89]). In the first section, the description of the approximate polynomial preconditioning technique is given. In section two, we discuss the minimax approximate polynomial applied to the system Eq.(8.0.1), and its convergence behavior. Finally, in section three, least squares polynomial preconditioning is considered.

## 8.1 The Approximate Polynomial Preconditioning

An important practical consideration in the PCG algorithm is that the preconditioning step should be easily computed. The computational work for each basic CG iteration essentially consists of one matrix by vector multiplication, two vector inner products, and three vector additions. For the polynomial preconditioning, the amount of work due to the preconditioning is about one matrix by vector multiplication for each power that appears in the polynomial. Consequently, the computational work per CG iteration may be dominated by the cost required for the preconditioning step when the degree $m$ of the preconditioning polynomial is very high. Thus it is important to use a preconditioning operator that requires a small amount of computational work. Wong [Wong88] observed that by using a preconditioning operator based on another matrix, a more efficient algorithm can be obtained. In this method the preconditioning operator $M^{-1}$ is chosen as a polynomial of another matrix, *the approximate matrix*, that approximates the original matrix, but has fewer non-zero entries. By using such preconditionings, the computational work in the preconditioning step 4 is significantly reduced, thus speeding up the overall PCG process. The success of this technique has been demonstrated in [Wong88] and [WoJi89].

When approximate polynomial preconditionings are applied to certain specially structured systems like Eq.(8.0.1), still better results can be achieved. We will propose a polynomial preconditioning operator that not only saves some computational work at each iteration, but also requires fewer iterations for convergence.

The system Eq.(8.0.1) is more difficult to solve than the system Eq.(7.1.1) for the following two reasons. First, Eq.(8.0.1) is worse conditioned than Eq.(7.1.1), because the former has a larger condition number. In fact, $\kappa(A^2) = \kappa(A)^2$ . Second, when the CG method is applied to Eq.(8.0.1), two matrix multiplications have to be carried out at each iteration, since in general one should avoid forming $A^2$ explicitly so that $A^2$ is computed by two applications of $A$ . When a traditional polynomial preconditioning method as described in chapter 7 is used, i.e. $M^{-1} = p_m(A^2)$ , two matrix multiplications are needed for each power of the polynomial $p_m$ . Therefore, for a fixed $m$ , each preconditioning step for the Eq.(8.0.1) is twice as expensive as for the Eq.(7.1.1), not forgetting that more iteration steps are needed for Eq.(8.0.1) because of its larger condition number. We can therefore conclude that the traditional polynomial preconditioning is not an efficient method to solve the system Eq.(8.0.1).

The performance of the PCG method totally relies on the eigensystem of the preconditioned matrix $M^{-1}A$ , which in the above case is $A^2 p_m(A^2)$ . Because the eigenvalues of $A^2$ are uniquely determined by $A$ , it is reasonable to expect that a polynomial in $A$ can provide as much information about the eigenvalue distribution as a polynomial in $A^2$ would. We thus propose a polynomial preconditioning matrix given by

$$M^{-1}_{APP(m)} = p_m(A)\,, \qquad p_m \in \Pi_m\,. \tag{8.1.1}$$

This preconditioning operator is referred to as the *approximate polynomial preconditioning operator*, abbreviated to APP operator.

An immediate advantage of this operator is that the computational cost at each preconditioning step is reduced by half, because only one matrix multiplication is needed for each power of $p_m$ . In the following sections, we will show that the convergence rate of this preconditioning is also more favorable. In fact, by properly choosing $p_m$ , we can show that the convergence rate by using Eq.(8.1.1) is comparable to the convergence rate of the PCG applied to the *better* conditioned system Eq.(7.1.1).

## 8.2 The Minimax Polynomial

As mentioned before, to select a polynomial that is best suitable for the PCG algorithm is a difficult task; it is in fact practically impossible without *a priori* knowledge of the solution. In this section, we will consider the polynomial that minimizes the condition number of the preconditioned system, i.e., the minimax polynomial.

When the preconditioning operator is given by Eq.(8.1.1), the resulting system is $M^{-1}_{APP(m)}A^2 = A^2 p_m(A), \quad p_m \in \Pi_m$ . Thus our problem is to find a polynomial that minimizes the condition number

$$\kappa(\lambda^2 p_m(\lambda)) = \frac{\max\limits_{\lambda \in [\lambda_1, \lambda_n]} \lambda^2 p_m(\lambda)}{\min\limits_{\lambda \in [\lambda_1, \lambda_n]} \lambda^2 p_m(\lambda)}, \tag{8.2.1}$$

where $0 < \lambda_1 \leq \ldots \leq \lambda_n$ are the eigenvalues of $A$ . Note that this polynomial is different from the one given in Eq.(7.2.5). The following notations will be used.

8.2.1 DEFINITION. *Let*

$$\Pi^{1,0}_{m+2} = \{ \; p_{m+2} \in \Pi_{m+2} \; | \; p_{m+2}(0) = 1, \; p'_{m+2}(0) = 0 \; \}.$$

*We denote by* $R_{m+2}$ *the unique polynomial in* $\Pi^{1,0}_{m+2}$ *such that*

$$\max_{\lambda \in [\lambda_1, \lambda_n]} |R_{m+2}(\lambda)| = \min_{p_{m+2} \in \Pi^{1,0}_{m+2}} \; \max_{\lambda \in [\lambda_1, \lambda_n]} |p_{m+2}(\lambda)|.$$

*We also define*

$$\theta_m = \max_{\lambda \in [\lambda_1, \lambda_n]} |R_{m+2}(\lambda)|.$$

The following lemma is a counterpart of Thm.7.2.3.

8.2.2 LEMMA. *Let* $R_{m+2}$ *be defined above, and* $q_m = \dfrac{1 - R_{m+2}}{\lambda^2}$ . *Then the condition number Eq.(8.2.1) satisfies*

$$\kappa(\lambda^2 q_m) \leq \kappa(\lambda^2 p_m), \qquad \text{for all} \quad p_m \in \Pi_m .$$

PROOF: Let $R_{m+2}, \theta_m$ be defined above. It can be shown (see Thm.9.2.1(3)) that $\theta_m < 1$ . By the definition of $q_m$ , we have $\kappa(\lambda^2 q_m) = \kappa(1 - R_{m+2}(\lambda)) = \dfrac{1 + \theta_m}{1 - \theta_m}$ . For any given $\tilde{p}_m \in \Pi_m$ , let

$$\Omega = \max_{\lambda \in [\lambda_1, \lambda_n]} \lambda^2 \tilde{p}_m(\lambda), \quad \text{and} \quad \omega = \min_{\lambda \in [\lambda_1, \lambda_n]} \lambda^2 \tilde{p}_m(\lambda).$$

Then we must show that

$$\frac{1+\theta_m}{1-\theta_m} \le \frac{\Omega}{\omega},$$

or equivalently,

$$\theta_m \le \frac{\Omega-\omega}{\Omega+\omega}.$$

To this end, we recall from the definition of $R_{m+2}$ that

$$\theta_m = \max_{\lambda\in[\lambda_1,\lambda_n]} |R_{m+2}(\lambda)| \le \max_{\lambda\in[\lambda_1,\lambda_n]} |p_{m+2}(\lambda)|$$

for any $p_{m+2} \in \Pi^{1,0}_{m+2}$ . In particular, this holds if $p_{m+2} \in \Pi^{1,0}_{m+2}$ is given by

$$p_{m+2} = 1 - \frac{2\lambda^2 \tilde{p}_m(\lambda)}{\Omega+\omega}.$$

For this $p_{m+2}$ , $\max_{\lambda\in[\lambda_1,\lambda_n]} |p_{m+2}(\lambda)| = \frac{\Omega-\omega}{\Omega+\omega}$. This completes the proof. □

Thus the polynomial which minimizes the condition number of the preconditioned system $M^{-1}_{APP(m)}A^2$ is $q_m$ given by the above lemma. Lem.8.2.2 also shows that $q_m \in \Pi_m$ is the minimum value of the expression

$$\max_{\lambda\in[\lambda_1,\lambda_n]} |1-\lambda^2 q_m(\lambda)|.$$

In this sense,

$$M^{-1}_{APP(m)} = q_m(A) \tag{8.2.2}$$

is the best approximation to the inverse matrix of $A^2$ among all polynomials of degree at most $m$ . In the following, we will study the convergence factor by using Eq.(8.2.2).

We denote by $\nu_{APP(m)}$ the convergence factor of PCG preconditioned by Eq.(8.2.2), i.e.,

$$\nu_{APP(m)} = \frac{\sqrt{\kappa(M^{-1}A^2)}-1}{\sqrt{\kappa(M^{-1}A^2)}+1}, \tag{8.2.3}$$

where $M^{-1} = M^{-1}_{APP(m)}$ . In the proof of Lem.8.2.2, we have shown that

$$\kappa(\lambda^2 q_m) = \frac{1+\theta_m}{1-\theta_m}.$$

The above expression is an upper bound on the condition number $\kappa(M^{-1}A^2)$ as shown in chapter 7. Therefore

$$\kappa(M^{-1}_{APP(m)}A^2) \le \kappa(\lambda^2 q_m) = \frac{1+\theta_m}{1-\theta_m}. \tag{8.2.4}$$

Combining Eq.(8.2.3) and Eq.(8.2.4), we obtain an upper bound of $\nu_{APP(m)}$

$$\nu_{APP(m)} \le \frac{\theta_m}{1+\sqrt{1-\theta_m^2}} \le \theta_m. \tag{8.2.5}$$

Some estimates for $\theta_m$ are given in the following lemmas.

8.2.3 LEMMA. *For* $m = 0, 1, 2, \dots$ *, the following estimates hold:*

$$\theta_m \le \begin{cases} \dfrac{1+2(m+1)^2}{T_{m+1}(\frac{\kappa+1}{\kappa-1})} \\ \\ \dfrac{1+\sqrt{\kappa}(m+1)}{T_{m+1}(\frac{\kappa+1}{\kappa-1})}, \end{cases} \tag{8.2.6}$$

*where* $\kappa = \kappa(A) = \lambda_n/\lambda_1$ *, and* $T_{m+1}$ *is the Chebyshev polynomial of degree* $m+1$ .

PROOF: Define $\tilde{T}_{m+1}(\lambda)$ by $\tilde{T}_{m+1}(\lambda) = T_{m+1}(\sigma(\lambda))$ , where

$$\sigma(\lambda) = -1 + 2\frac{\lambda-\lambda_1}{\lambda_n-\lambda_1}.$$

If we define

$$p_{m+2}(\lambda) = \left(1 - \frac{\tilde{T}'_{m+1}(0)}{\tilde{T}_{m+1}(0)}\lambda\right)\frac{\tilde{T}_{m+1}(\lambda)}{\tilde{T}_{m+1}(0)},$$

then $p_{m+2} \in \Pi^{1,0}_{m+2}$ . By the minimizing property of $R_{m+2}$ in Def.8.2.1, one has

$$\theta_m = \max_{\lambda\in[\lambda_1,\lambda_n]} |R_{m+2}(\lambda)| \le \max_{\lambda\in[\lambda_1,\lambda_n]} |p_{m+2}(\lambda)|. \tag{8.2.7}$$

An estimate for $|p_{m+2}|$ can be obtained by using the expression for the Chebyshev polynomial

$$T_{m+1}(\lambda) = \tfrac{1}{2}[(\lambda + \sqrt{\lambda^2 - 1})^{m+1} + (\lambda - \sqrt{\lambda^2 - 1})^{m+1}].$$

Direct calculations show that

$$\left|\frac{\tilde{T}'_{m+1}(0)}{\tilde{T}_{m+1}(0)}\right| = \frac{m+1}{\sqrt{\lambda_1 \lambda_n}} \frac{1 - \left(\dfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{m+1}}{1 + \left(\dfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{m+1}}.$$

By the definition of $p_{m+2}$, it is easy to show that

$$|p_{m+2}| \leq \left\{ 1 + \sqrt{\kappa}(m+1) \frac{1 - \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{m+1}}{1 + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{m+1}} \right\} T^{-1}_{m+1}\left(\frac{\kappa+1}{\kappa-1}\right).$$

Then Eq.(8.2.6) follows by combining the above and Eq.(8.2.7). □

When $\kappa$ is large compared to the degree $m$ of the preconditioning polynomial, the first bound in Eq.(8.2.6) should be used. Otherwise the second should be used. The following lemma gives an asymptotic estimate of $\theta_m$ when $m$ is very large.

8.2.4. LEMMA [Bern26]. *The following equality holds:*

$$\lim_{m \to \infty} \theta_m \cdot \left\{ \frac{8\sqrt{\kappa}}{\kappa - 1} \frac{m+2}{T_{m+1}(\frac{\kappa+1}{\kappa-1})} \right\}^{-1} = 1,$$

*that is*

$$\theta_m \approx \frac{8\sqrt{\kappa}}{\kappa - 1} \frac{m+2}{T_{m+1}(\frac{\kappa+1}{\kappa-1})}, \quad \text{as} \quad m \to \infty.$$

PROOF: See [Bern26]. □

Lem.8.2.3 gives a bound that is too pessimistic. In fact, the right hand side of Eq.(8.2.6) cannot be bounded by 1 for all $m$, even though it tends to 0 as $m \to \infty$. Lem.8.2.4 provides an accurate estimate only when $m$ is very large. However, in practice, the degree of the preconditioning polynomial is not particularly large. For completeness, we provide the following lemma, the proof of which has to be deferred until Chapter 9.

8.2.5 LEMMA. *There exist* $0 < \alpha < \beta$, *dependent on* $m, \lambda_1$ *and* $\lambda_n$ *only, such that*

$$\theta_m = \left\{ \cosh\{(m+1)\int_0^{\lambda_1} \frac{\lambda\, d\lambda}{\sqrt{(\lambda+\alpha)(\lambda+\beta)(\lambda_1-\lambda)(\lambda_n-\lambda)}}\} \right\}^{-1} .$$

PROOF: This follows from Lem.9.2.4 of Chapter 9. ☐

The significance of the above lemma is that once $m, \lambda_1$ and $\lambda_n$ are given, one can compute numerically a very accurate estimate for the convergence factor, hence the number of iterations can be estimated.

8.2.6 LEMMA. *Let* $\theta_{2m}$ *be the maximum value of* $|R_{2m+2}|$ *in* $[\lambda_1, \lambda_n]$. *Then*

$$\theta_{2m} < \frac{1}{T_{m+1}\left(\frac{\kappa^2+1}{\kappa^2-1}\right)} .$$

PROOF: Let

$$p_{2m+2} = T_{m+1}\left(\frac{2\lambda^2 - \lambda_n^2 - \lambda_1^2}{\lambda_n^2 - \lambda_1^2}\right) T_{m+1}^{-1}\left(\frac{\kappa^2+1}{\kappa^2-1}\right) .$$

Then $p_{2m+2} \in \Pi_{2m+2}^{1,0}$. By the minimizing property of $R_{2m+2}$, one has

$$\theta_{2m} = \max_{\lambda \in [\lambda_1, \lambda_n]} |R_{2m+2}| < \max_{\lambda \in [\lambda_1, \lambda_n]} |p_{2m+2}| ,$$

and the lemma follows from properties of the Chebyshev polynomial. ☐

To compare the convergence of APP preconditioning with that of CG applied to Eq.(7.1.1), we will give the estimate in terms of $\nu$ .

8.2.7 THEOREM. *With* $M^{-1}_{APP(m)} = q_m(A)$ , *the APP preconditioning yields a convergence factor*

$$\nu_{APP(m)} \le \begin{cases} 2(1 + 2(m+1)^2)\nu^{m+1} \\ 2(1 + \sqrt{\kappa}(m+1))\nu^{m+1} , \end{cases} \tag{8.2.8}$$

*where* $\nu = \dfrac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$ *and* $\kappa = \lambda_n / \lambda_1$ .

PROOF: Since $\nu = \dfrac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$ , the property of the Chebyshev polynomial leads to

$$T_{m+1}\left(\frac{\kappa+1}{\kappa-1}\right) = \frac{1 + \nu^{2m+2}}{2\nu^{m+1}} \ge \frac{1}{2\nu^{m+1}} .$$

Then the result follows from Eq.(8.2.5) and Eq.(8.2.6). □

Since $\nu < 1$ , the term $\nu^{m+1}$ dominates the right hand side of the Eq.(8.2.8) when $m$ is large. Thus the estimate for $\nu_{APP(m)}$ is comparable to that of Thm.7.2.4 or Thm.7.2.5. This leads to the important conclusion that *using the APP preconditioning* $M^{-1}_{APP(m)} = q_m(A)$ *for the system* $A^2u = b$ *yields a convergence rate that is comparable to using the minimax or the best least squares polynomial preconditioning of same degree for the better conditioned system* $Au = b$ .

Next, we will compare the APP preconditioning with the traditional polynomial preconditioning, by which we mean the methods of chapter 7 applied directly to the system Eq.(8.0.1). For convenience, the latter method will be called *TPP preconditioning*. We shall show that APP preconditioning is more efficient than TPP preconditioning.

Let $\nu_{TPP(m)}$ be the convergence factor of the TPP preconditioning $M^{-1} =$

$p_m(A^2)$ where $p_m$ is the minimax polynomial of Thm.7.2.3. Then it follows from Thm.7.2.4 that

$$\nu_{TPP(m)} = \nu_{A^2}^{m+1}\,, \qquad \text{where} \quad \nu_{A^2} = \frac{\sqrt{\kappa(A^2)}-1}{\sqrt{\kappa(A^2)}+1}\,. \tag{8.2.9}$$

8.2.8. COROLLARY. *The convergence factors of the APP and TPP satisfy*

$$\lim_{m\to\infty} \frac{\nu_{APP(m)}}{\nu_{TPP(m)}} = 0\,.$$

PROOF: By using the first estimate in the Eq.(8.2.8), and Eq.(8.2.9), we have the inequality

$$\frac{\nu_{APP(m)}}{\nu_{TPP(m)}} \le C(m+2)^2 \left(\frac{\nu}{\nu_{A^2}}\right)^{m+1},$$

where $C$ is a constant independent of $m$. We claim that the ratio on the right hand side of the above is smaller than 1. In fact,

$$\nu_{A^2} = \frac{\sqrt{\kappa(A^2)}-1}{\sqrt{\kappa(A^2)}+1} = \frac{\kappa(A)-1}{\kappa(A)+1} > \frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1} = \nu\,,$$

because $\kappa > 1$. Therefore $\dfrac{\nu}{\nu_{A^2}} < 1$, and the corollary follows. □

The above corollary shows that there exists an $m_0$ such that if $m > m_0$ then $\nu_{APP(m)} < \nu_{TPP(m)}$, i.e., APP preconditioning has a faster convergence rate than TPP preconditioning. The value $m_0$ depends on the value of the condition number $\kappa(A)$. From Cor.8.2.8, it may seem that $m_0$ must be very large if $\kappa(A)$ is large. However, it must be pointed out that Cor.8.2.8 is rather pessimistic, because the bound for $\nu_{APP(m)}$ given in Thm.8.2.7 is too crude while that for $\nu_{TPP(m)}$ as given in Eq.(8.2.9) is relatively sharp. The numerical experiments to be reported in chapter 10 show that APP preconditioning yields a faster convergence than TPP for any value of $m$.

Even though we are not be able to prove that $\nu_{APP(m)} < \nu_{TPP(m)}$ for all $m$, we will, however, show that, for any $m$, APP preconditioning is more efficient than TPP preconditioning, in the sense that APP requires less computational work than TPP to yield the same accuracy of approximations of the solution. For a fixed degree $m$ of the preconditioning polynomials, computing $M^{-1}_{TPP(m)}r = p_m(A^2)r$ is twice as expensive as computing $M^{-1}_{APP(m)}r = q_m(A)r$. Therefore $2m$ term APP preconditioning $M^{-1}_{APP(2m)}$ requires the same amount of computation as $m$ term TPP preconditioning at each iteration of the PCG algorithm. Thus to determine which preconditioning is more efficient, we should compare the convergence factors of $M^{-1}_{APP(2m)}$ and $M^{-1}_{TPP(m)}$.

8.2.9. **THEOREM.** *For* $m = 0, 1, \ldots$, *we have*

$$\nu_{APP(2m)} < \nu_{TPP(m)}.$$

PROOF: When $2m$ is used in Eq.(8.2.5), the result is

$$\nu_{APP(2m)} \le \frac{\theta_{2m}}{1 + \sqrt{1 - \theta_{2m}^2}}.$$

From Lem.8.2.6, we have

$$\theta_{2m} < T^{-1}_{m+1}\left(\frac{\kappa^2 + 1}{\kappa^2 - 1}\right) = \frac{2\nu_{A^2}^{m+1}}{1 + \nu_{A^2}^{2m+2}}.$$

Combining the above two inequalities, we find that

$$\nu_{APP(2m)} < \nu_{A^2}^{m+1} = \nu_{TPP(m)}.$$

□

This theorem shows that the $2m$ term APP converges faster than the $m$ term TPP while they both require the same amount of computational work at each iteration step.

## 8.3 The Least Squares Polynomial

In chapter 7, we have pointed out the reasons for using the least squares polynomial in polynomial preconditionings. However, there is an extra advantage of the least squares polynomial when the approximate polynomial preconditioning technique is used. We have seen from the previous section that the minimax polynomial is determined by $R_{m+2}$ of Def.8.1.1. This polynomial is however very difficult to compute, in contrast to the traditional polynomial preconditioning where the minimax polynomial can be calculated from the Chebyshev polynomial. On the other hand, the least squares polynomial is alway easy to compute.

The least squares polynomial preconditioning operator is given by

$$\tilde{M}^{-1} = \tilde{q}_m(A),$$

where $\tilde{q}_m \in \Pi_m$ as determined by

$$\int_{\lambda_1}^{\lambda_n} \left(1 - \lambda^2 \tilde{q}_m(\lambda)\right)^2 d\lambda = \min_{p_m \in \Pi_m} \int_{\lambda_1}^{\lambda_n} \left(1 - \lambda^2 p_m(\lambda)\right)^2 d\lambda. \qquad (8.3.1)$$

If we define the residual polynomial by

$$\tilde{R}_{m+2}(\lambda) = 1 - \lambda^2 \tilde{q}_m(\lambda),$$

then $\tilde{R}_{m+2} \in \Pi^{1,0}_{m+2}$ and Eq.(8.3.1) becomes

$$\int_{\lambda_1}^{\lambda_n} \left|\tilde{R}_{m+2}(\lambda)\right|^2 d\lambda = \min_{p_{m+2} \in \Pi^{1,0}_{m+2}} \int_{\lambda_1}^{\lambda_n} \left|p_{m+2}(\lambda)\right|^2 d\lambda. \qquad (8.3.2)$$

In the following, we proceed in much the same way as in the previous section. We define

$$\tilde{\theta}_m = \max_{\lambda \in [\lambda_1, \lambda_n]} |\tilde{R}_{m+2}(\lambda)|,$$

then the convergence factor $\tilde{\nu}_{APP(m)}$ by using $\tilde{M}^{-1}_{APP(m)} = \tilde{q}_m(A)$ as preconditioning operator satisfies the estimate

$$\tilde{\nu}_{APP(m)} \leq \tilde{\theta}_m . \tag{8.3.3}$$

We will find a bound for $\tilde{\theta}_m$ .

8.3.1 LEMMA [JMP83]. *Let* $p_m \in \Pi_m$ , *and* $x \in [\lambda_1, \lambda_n]$ . *Then*

$$|p_m(x)|^2 \leq \frac{m^2+2m}{\lambda_n - \lambda_1} \int_{\lambda_1}^{\lambda_n} |p_m(\lambda)|^2 d\lambda .$$

PROOF: Johnson *et al.* [JMP83] have shown this for $\lambda_1 = -1,\ \lambda_n = 1$ . The above lemma holds if one uses the transformation $t = -1 + 2\dfrac{\lambda - \lambda_1}{\lambda_n - \lambda_1}$ in the result of [JMP83]. □

From this lemma, we can prove

8.3.2 THEOREM. *If the best least squares polynomial defined by Eq.(8.3.1) is used in APP, its convergence factor satisfies*

$$\tilde{\nu}_{APP(m)} \leq 6(m+2)^3 \nu^{m+1} .$$

PROOF: From Lem.8.3.1, we find

$$\tilde{\theta}_m^2 = \max_{\lambda \in [\lambda_1, \lambda_n]} |\tilde{R}_{m+2}(\lambda)|^2 \leq \frac{(m+2)^2 + 2(m+2)}{\lambda_n - \lambda_1} \int_{\lambda_1}^{\lambda_n} \left|\tilde{R}_{m+2}(\lambda)\right|^2 d\lambda .$$

Let $R_{m+2}$ be defined as in Def.8.2.1. Then by the minimization properties of $\tilde{R}_{m+2}$ , the following inequality holds:

$$\int_{\lambda_1}^{\lambda_n} \left|\tilde{R}_{m+2}(\lambda)\right|^2 d\lambda \leq \int_{\lambda_1}^{\lambda_n} \left|R_{m+2}(\lambda)\right|^2 d\lambda \leq \theta_m^2 (\lambda_n - \lambda_1) ,$$

where $\theta_m$ is the maximum of $|R_{m+2}|$ in the interval $[\lambda_1, \lambda_n]$. The above two inequalities, when combined, yield

$$\tilde{\theta}_m^2 \le \left((m+2)^2 + 2(m+2)\right)\theta_m^2 \le 2(m+2)^2\theta_m^2 . \tag{8.3.4}$$

The first bound of $\theta_m$ in Eq.(8.2.6) yields

$$\theta_m \le \frac{1+2(m+1)^2}{T_{m+1}\left(\frac{\kappa+1}{\kappa-1}\right)} \le \frac{2(m+2)^2}{T_{m+1}\left(\frac{\kappa+1}{\kappa-1}\right)} .$$

When this is inserted into Eq.(8.3.4), the result is

$$\tilde{\theta}_m \le \frac{3(m+2)^3}{T_{m+1}\left(\frac{\kappa+1}{\kappa-1}\right)} .$$

The theorem follows from $T_{m+1}^{-1}(\frac{\kappa+1}{\kappa-1}) \le 2\nu^{m+1}$ and Eq.(8.3.3). ☐

By comparing this theorem with Thm.7.2.5, we observe that the bound for $\tilde{\nu}_{APP(m)}$ compares quite favorably with that for $\tilde{\nu}$, which is the convergence factor of the best least squares polynomial preconditioning applied to the system $Au = b$. We therefore conclude that *the least squares APP preconditioning* $\tilde{M}_{APP}^{-1}$ *is also more efficient than the TPP preconditioning.*

To conclude this section, we dwell briefly on the computation of the least squares polynomial. The coefficients of this polynomial can be obtained by solving a linear normal equation. However, if the value of $m$ is large, the normal equation becomes ill-conditioned, resulting in a loss of accuracy. An alternative is to use a basis of orthogonal polynomials.

Define an inner product by

$$\langle p, q\rangle_w = \int_{\lambda_1}^{\lambda_n} (1-\lambda)^4 p(\lambda) q(\lambda) d\lambda .$$

Then an orthogonal basis $\{\psi_j\}$ of $\Pi_m$ with respect to $\langle\cdot,\cdot\rangle_w$ can be found by a three term recursive formula. In fact, it can be shown that

$$\psi_j(\lambda) = \frac{a_j}{(\lambda-1)^4}\frac{d^j}{d\lambda^j}(\lambda-1)^{j+4}(\lambda+1)^j ,$$

where $a_j$ is some constant. When the best least squares polynomial $\tilde{q}_m$ of Eq.(8.3.1) is expressed in terms of this base, the result is

$$\tilde{q}_m = \sum_{j=0}^{k} c_j\psi_j , \qquad c_j = \frac{\langle f, \psi_j\rangle_w}{\langle \psi_j, \psi_j\rangle_w} ,$$

where $f(\lambda) = \dfrac{1}{(1-\lambda)^2}$ . The $c_j$ can be calculated analytically.

# CHAPTER 9

# COMPUTATION OF THE MINIMAX POLYNOMIAL

We have shown that finding the polynomial that minimizes the condition number of the preconditioned system $M_{APP}^{-1}A^2$ is tantamount to finding the constrained minimax polynomial $R_{m+2}$ of Def.8.2.1. This polynomial is however not explicitly known; numerical methods must be used to compute it. The theme of this chapter is to develop an efficient algorithm to compute the coefficients of the minimax polynomial required by the APP preconditioning. In section one, we present a problem where the polynomial $R_{m+2}$ is also used, and review the existing algorithm for computing it. In section two, the properties of $R_{m+2}$ are studied. Finally, in section three, we present our new algorithm to calculate the coefficients of the minimax polynomial.

## 9.1 Introduction

In the previous chapter, the constrained minimax polynomial $R_{m+2}$ was introduced in order to construct the approximate polynomial that minimizes the condition number of the preconditioned system. The constraint $R'_{m+2}(0) = 0$ results from the special structure of the model problem $A^2u = b$, because $R_{m+2}$ is the residual polynomial of the preconditioned system $M^{-1}A$, i.e., $R_{m+2}(\lambda) = 1 - \lambda^2 q_m(\lambda)$. Therefore it is perhaps hard to imagine that this polynomial should have any relation with the preconditioning polynomials of any other system, say $Au = b$. However, surprisingly enough, the method developed in chapter 8 of this thesis is not the only application of $R_{m+2}$; $R_{m+2}$ also appears in another polynomial preconditioning algorithm for the solution of the system $Au = b$ [Ashl87].

Recall from chapter 7 that the PCG algorithm computes a sequence of approximations that minimizes the $A$-norm of the error $\|e_j\|_A$ at each iteration step. However, it would be more desirable to obtain a preconditioning which minimizes the Euclidean norm of the error, $\|e_j\|_2$. At first glance, such an algorithm seems impossible, because it requires the calculation of $\langle e_j, d_j \rangle$ at step 1 of the PCG algorithm, where $e_j$ (the error) is unknown. But as Manteuffel observed [Ashl87], if the preconditioning operator is given by

$$M^{-1} = Ap_m(A), \tag{9.1.1}$$

then $\langle e_j, d_j \rangle$ can be computed by the formula

$$\langle e_j, d_j \rangle = \langle e_j, M^{-1} r_j \rangle = \langle e_j, Ap_m(A) r_j \rangle = \langle r_j, p_m(A) r_j \rangle ,$$

where $r_j$ is the residual. Therefore when Eq.(9.1.1) is used as a preconditioning operator, it is practical to minimize the Euclidean norm of the error. To find such a polynomial which minimizes the condition number of the preconditioned system $M^{-1}A = A^2 p_m(A)$, $p_m$ must be the minimum value in $\Pi_m$ of the expression

$$\max_{\lambda \in [\lambda_1, \lambda_n]} |1 - \lambda^2 p_m(\lambda)| .$$

This is precisely the constrained minimax problem of Def.8.2.1.

Unfortunately, a difficulty arises when one wants to calculate $R_{m+2}$, because, unlike the Chebyshev polynomial, $R_{m+2}$ is not explicitly known, and may not be generated by a three-term recursion. The method used in the past to compute $R_{m+2}$ numerically is the *Remez algorithm*. The Remez algorithm (see e.g. [Chen66]) determines the minimax polynomial by solving a succession of discrete minimax approximation problems; the discrete problem is solved on a discrete set called the *trial alternant*, this alternant is then updated and the process repeated. Thus each iteration consists of (1) solving a discrete minimax problem, which is equivalent to solving a linear system, and (2) selecting a new trial alternant, which includes locating all roots of a polynomial (see description in [Ashl87]). The Remez algorithm is necessarily complicated, because it is versatile enough to apply to general minimax polynomial approximation problems. The method is powerful, yet inefficient when applied to compute the polynomial $R_{m+2}$ .

By employing some properties of $R_{m+2}$ , a more efficient algorithm can be designed. In our new algorithm, to find the coefficients of the minimax preconditioning polynomial $q_m$ , only the solution of one $(m+1)\times(m+1)$ linear system and one pair of roots of two nonlinear scalar equations need to be computed.

## 9.2 Characterization of $R_{m+2}$

In this and the next section, the independent variable of polynomials will be denoted by $x$ instead of $\lambda$ . The interval of interest is denoted by $[a,b]$ , where it is always assumed that $0 < a < b$ .

We let

$$p_{m+2}(x) = \sum_{j=0}^{m} c_j x^{j+2}$$

be the *generalized polynomial* of the Chebyshev system (see [Chen66, p72]):

$$x^2, \quad x^3, \quad \dots, \quad x^{m+2}.$$

If we can find the best approximation $p_{m+2}$ of 1 over the interval $[a,b]$ , then $R_{m+2}$ is its error function, where $R_{m+2} \in \Pi^{1,0}_{m+2}$ is given by

$$\max_{x\in[a,b]} |R_{m+2}(x)| = \min_{p_{m+2}\in\Pi^{1,0}_{m+2}} \max_{x\in[a,b]} |p_{m+2}(x)|. \qquad (9.2.1)$$

9.2.1 THEOREM. *For* $m \geq 0$ *, let* $R_{m+2}$ *be defined as above and*

$$\theta_m = \max_{x \in [a,b]} |R_{m+2}(x)| .$$

*Then*

*(1) there are exactly* $m+2$ *points* $a = x_0 < x_1 < \ldots < x_{m+1} = b$ *, such that*

$$R_{m+2}(x_j) = -R_{m+2}(x_{j+1}) = (-1)^j \theta_m , \quad j = 0, 1, \ldots, m+1 .$$

*(2)* $x = 0$ *is a local maximum of* $R_{m+2}$ *, and it is the only extreme point of* $R_{m+2}$ *outside the interval* $[a,b]$ *.*

*(3)* $\theta_m < 1$ *.*

*(4) there are* $\alpha, \beta : 0 < \alpha < \beta$ *such that* $R_{m+2}(-\alpha) = -R_{m+2}(-\beta) = \theta_m$ *.*

PROOF:

(1) The existence of $x_j$ is a direct consequence of the Alternation Theorem. See [Chen66,p75].

(2) This follows from (1) by counting the zeros of $R'_{m+2}$ .

(3) By (2) $R_{m+2}$ is strictly decreasing on the interval $[0,a]$ , therefore $\theta_m = R_{m+2}(a) < R_{m+2}(0) = 1$ .

(4) $R_{m+2}$ is strictly increasing on the interval $(-\infty, 0]$ and

$$\lim_{x \to -\infty} R_{m+2}(x) = -\infty, \quad \lim_{x \to 0} R_{m+2}(x) = 1 > \theta_m .$$

This guarantees the existence of $\alpha$, $\beta$ . ☐

For an illustration of $R_{m+2}$ , see Fig.9.3.1. The set $\{x_0, x_1, \ldots, x_m, x_{m+1}\}$ is called the *alternant* of $R_{m+2}$ .

9.2.2 THEOREM. $R_{m+2}$ *satisfies the differential equation*

$$(x+\alpha)(x+\beta)(x-a)(x-b)p'^2 = (m+2)^2 x^2 (p^2 - \theta_m^2) , \qquad (9.2.2)$$

*where* $\alpha, \beta$ *are given by Thm.9.2.1.*

PROOF: By Thm.9.2.1, $R'^2_{m+2}$ has double zeros at $0, x_1, x_2, \ldots, x_m$ . Also $R^2_{m+2} - \theta^2_m$ has double zeros at $x_1, \ldots, x_m$ and simple zeros at $a, b, \alpha, \beta$ . Therefore both sides of Eq.(8.2.2), which are polynomials of the same degree, have the same zeros, counted with multiplicities. By comparing the leading coefficients, one shows that Eq.(9.2.2) holds. ☐

The differential equation Eq.(9.2.2) is nonlinear. However, it can be linearized as follows.

9.2.3 LEMMA. *Let* $q$ *be a function on* $[a, b]$ *. If* $q \neq$ *const. is a solution of Eq.(9.2.2), then it is also a solution of*

$$xf(x)p'' - \frac{1}{2}\big(f(x) - \frac{1}{2}xf'(x)\big)p' - (m+2)^2 x^3 p = 0\,, \qquad (9.2.3)$$

*where* $f(x) = (x+\alpha)(x+\beta)(x-a)(x-b)$ .

*Conversely, if* $q$ *is a solution of Eq.(9.2.3), it is also a solution of Eq.(9.2.2) where* $\theta_m = \max_{x \in [a,b]} |q(x)|$ .

PROOF: Assume $q \neq$ const. is a solution of Eq.(9.2.2). Differentiating both sides of Eq.(9.2.2) and dividing through by $q'$ , we obtain Eq.(9.2.3). If the process is reversed, we see that the solution of Eq.(9.2.3) satisfies Eq.(9.2.2) for some $\theta^2_m \in \mathbb{R}$ . By comparing the sign on both sides, we conclude that $\theta^2_m \geq 0$ and $\theta_m$ is the maximum of $|q|$ on $[a, b]$ . ☐

9.2.4 LEMMA. $R_{m+2}$ *is given by*

$$R_{m+2}(x) = \begin{cases} \theta_m \cos\left\{(m+2)\int_a^x \frac{t\,dt}{\sqrt{(t+\alpha)(t+\beta)(t-a)(b-t)}}\right\}, & x \in [a, b] \\ \theta_m \cosh\left\{(m+2)\int_{-\alpha}^x \frac{t\,dt}{\sqrt{(t+\alpha)(t+\beta)(a-t)(b-t)}}\right\}, & x \in [-\alpha, a], \end{cases}$$

*for some* $\theta_m$, $\alpha$, $\beta$ .

PROOF: From Thm.9.2.2, $R_{m+2}$ satisfies

$$\frac{p'^2}{p^2-\theta_m^2} = \frac{(m+2)^2x^2}{(x+\alpha)(x+\beta)(x-a)(x-b)},$$

and the lemma follows by integration. □

With the aid of this lemma, we can prove Lem.8.2.5 of Chapter 8. In fact, $\theta_m$ can be solved from the second expression of Lem.9.2.4 by letting $x = 0$ and using $R_{m+2}(0) = 1$ . But $\alpha$, $\beta$ remain undetermined. They can be determined by the following lemma.

9.2.5 LEMMA. *Let*

$$\begin{aligned} f^1(\alpha,\beta) &= \int_a^b \frac{tdt}{\sqrt{(t+\alpha)(t+\beta)(t-a)(b-t)}}, \\ f^2(\alpha,\beta) &= \int_0^\alpha \frac{tdt}{\sqrt{(\alpha-t)(\beta-t)(t+a)(t+b)}} \\ &\quad - \int_0^a \frac{tdt}{\sqrt{(t+\alpha)(t+\beta)(a-t)(b-t)}}. \end{aligned}$$

*Then*

*(1)* $\alpha$, $\beta$ *of Thm.9.2.2 satisty*

$$f^1(\alpha,\beta) = \frac{(m+1)\pi}{m+2} \tag{9.2.4a}$$

$$f^2(\alpha,\beta) = 0 . \tag{9.2.4b}$$

*(2) Eq.(9.2.4) has exactly one pair of solutions* $\alpha$, $\beta$ , *with* $0 < \alpha < \beta$ .

PROOF:

(1) We first show that Eq.(9.2.4) is satisfied. Since $R_{m+2}(b) = \pm\theta_m$ , it

follows from Lem.9.2.4 that

$$(m+2)\int_a^b \frac{tdt}{\sqrt{(t+\alpha)(t+\beta)(t-a)(b-t)}} = k\pi\,,$$

where $0 \le k \le m+2$ . By Thm.9.1.1, $R_{m+2}$ alternates exactly $m+2$ times on the interval $[a,b]$ . This leads to $k = m+1$ , which shows Eq.(9.2.4a). On the interval $[-\alpha, a]$ , $R_{m+2}(-\alpha) = R_{m+2}(a)$ . This implies, by the second formula in Lem.9.2.4,

$$\int_{-\alpha}^{a} \frac{tdt}{\sqrt{(t+\alpha)(t+\beta)(a-t)(b-t)}} = 0\,,$$

and Eq.(9.2.4b) follows.

(2) Each equation in Eq.(9.2.4) defines a curve in the $\alpha$, $\beta$ plane. We will show that these two curves have exactly one point of intersection in the region $0 < \alpha < \beta$ .

$f^1$ is a decreasing function of $\alpha$, $\beta$ . Direct calculation shows $f^1(0,0) = \pi > \frac{(m+1)\pi}{m+2}$ . Also $f^1(0,\beta) \to 0$ as $\beta \to \infty$ . This implies that there is a $\beta_0$ such that $(\alpha,\ \beta) = (0, \beta_0)$ is a solution of Eq.(9.2.4a). From the monotonicity of $f^1$ , it can be easily shown that for each $\beta \in [0, \beta_0]$ , Eq.(9.2.4a) defines a unique $\alpha = \alpha_1(\beta)$ . $\alpha_1(\beta)$ is a decreasing function of $\beta$ as shown in Fig.9.2.1.

By a change of variable, $f^2$ becomes

$$\begin{aligned} f^2(\alpha,\beta) = &\int_0^1 \frac{tdt}{\sqrt{(1-t)(\frac{\beta}{\alpha}-t)(t+\frac{a}{\alpha})(t+\frac{b}{\alpha})}} \\ &- \int_0^a \frac{tdt}{\sqrt{(t+\alpha)(t+\beta)(a-t)(b-t)}}\,. \end{aligned}$$

For a given $\beta > 0$ , $f^2(\alpha,\beta)$ increases monotonically from a negative number to $+\infty$ as $\alpha$ increases from $0$ to $\beta$ . Thus Eq.(9.2.4b) defines a smooth curve $\alpha = \alpha_2(\beta)$ in the region $0 < \alpha < \beta$ . It is also easy to show that $\lim\limits_{\beta\to 0} \alpha_2(\beta) =$

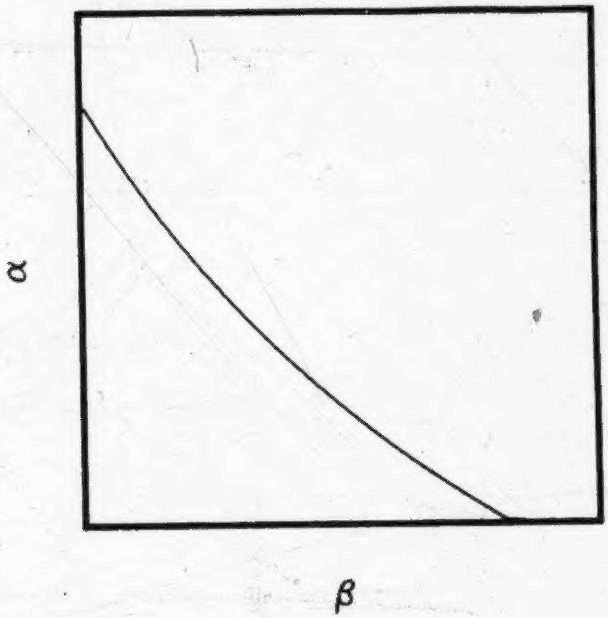

**Figure 9.2.1** The function $\alpha = \alpha_1(\beta)$ .

0, $\lim_{\beta\to\infty} \alpha_2(\beta) = \alpha_0 > 0$ . Now we show that $\alpha_2(\beta)$ is a monotonically increasing function of $\beta$ .

If $\alpha_2(\beta)$ were not a monotonic function of $\beta$ , there would be a $\beta > 0$ , such that $f^2_\alpha(\alpha_2(\beta),\beta) = \infty$ or $f^2_\beta(\alpha_2(\beta),\beta) = 0$ . But by taking derivatives, we find that $f^2_\alpha(\alpha_2(\beta),\beta)$ is finite for all $\beta > 0$ because $0 < \alpha_2(\beta) < \beta$ , and that $f^2_\beta(\alpha_2(\beta),\beta) \neq 0$ for all $\beta > 0$ . This contradiction shows that $\alpha_2(\beta)$ is monotonic. Then, it can only be increasing. $\alpha_2(\beta)$ is shown in Fig.9.2.2.

Therefore $\alpha = \alpha_1(\beta)$ and $\alpha = \alpha_2(\beta)$ have exactly one point in common in the region $0 < \alpha < \beta$ . This completes the proof. ☐

9.2.6 THEOREM. *Let* $\alpha$, $\beta$ *be the unique solution of Eq.(9.2.4). Then* $R_{m+2}$ *is the unique polynomial solution of Eq.(9.2.3) satisfying* $p(0) = 1$ .

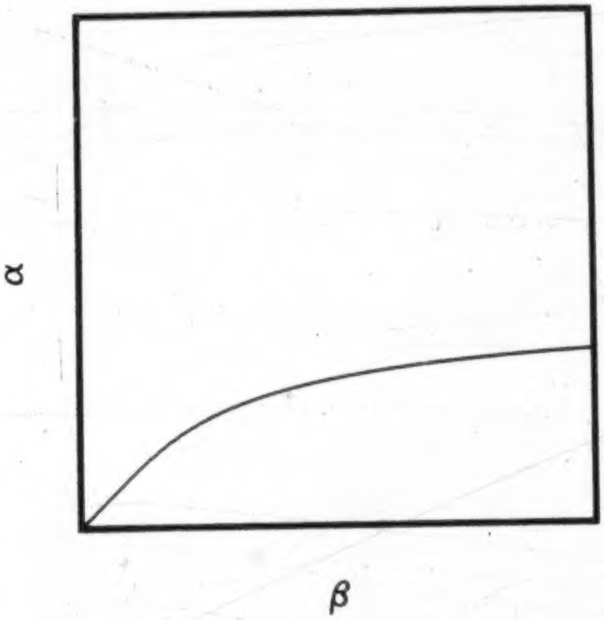

**Figure 9.2.2** The function $\alpha = \alpha_2(\beta)$ .

PROOF: That $R_{m+2}$ is a solution follows from Thm.9.2.2 and Lem.9.2.3. We now assume that $p$ is a polynomial satisfying Eq.(9.2.3) and $p(0) = 1$ . By Lem.9.2.3, $p$ is a solution of Eq.(9.2.2). Since $\alpha, \beta$ satisfy Eq.(9.2.4), $p$ satisfies the alternation property as stated in Thm.9.2.1(1). Again by the Alternation Theorem, $p = R_{m+2}$ . □

## 9.3 An Algorithm for the Computation of $R_{m+2}$

We have shown in the previous section that $R_{m+2}$ can be determined by solving the differential equation Eq.(9.2.3), which is a linear equation. Thus the original minimax problem is reduced to a *linear* problem after $\alpha$, $\beta$ are determined from Eq.(9.2.4).

To determine the coefficients of $q_m(x) = \sum_{j=0}^{m} c_j x^j$ , we substitute $R_{m+2}(x) = 1 + x^2 q_m(x)$ into Eq.(9.2.3). Equating the coefficients of like powers of $x$ , we obtain a linear system

$$Nc = h\,, \tag{9.3.1}$$

where $N = [N_{i,j}]$ is an $(m+3) \times (m+1)$ matrix, and $c$ is an $m+1$ -vector and $h$ an $m+3$ -vector.

The entries of $N$ are given by

$$\begin{aligned}
N_{1,1} &= -(\xi B - A\eta)\\
N_{1,2} &= 3A\xi\\
N_{2,1} &= 2(\xi + A - B\eta)\\
N_{2,2} &= -\frac{9}{2}(\xi B - A\eta)\\
N_{2,3} &= 8A\xi\\
N_{3,1} &= 3(\eta - b)\\
N_{3,2} &= 6(\xi + A - B\eta)\\
N_{3,3} &= -10(\xi B - A\eta)\\
N_{3,4} &= 15A\xi\,.
\end{aligned} \tag{9.3.2a}$$

For $j = 2, \ldots, m+1$ ,

$$\begin{aligned}
N_{j+2,j-1} &= (j - m + 2)(n + j + 2)\\
N_{j+2,j} &= \frac{(j+1)(2j+1)}{2}(\eta - B)\\
N_{j+2,j+1} &= (j+2)(j+1)(\xi + A - B\eta)H(m-j)\\
N_{j+2,j+2} &= -\frac{(2j+3)(j+3)}{2}(\xi B - A\eta)H(m-j-1)\\
N_{j+2,j+3} &= (j+2)(j+4)A\xi H(m-j-2)\,.
\end{aligned} \tag{9.3.2b}$$

where $\xi = \alpha\beta,\ \eta = \alpha + \beta,\ A = ab,\ B = a + b$ , and

$$H(j) = \begin{cases} 0, & j < 0 \\ 1, & j \geq 0 . \end{cases}$$

Finally,

$$N_{i,j} = 0, \qquad \text{otherwise} . \tag{9.3.2c}$$

On the right hand side, $h_2 = (m+2)^2,\ h_j = 0,\ j \neq 2$ .

9.3.1 LEMMA. *The rank of* $N$ *is* $m + 1$ *for all* $\alpha,\ \beta$ .

PROOF: We need to show that

$$Nc = 0 \tag{9.3.3}$$

implies $c = 0$ . Suppose that $c$ is a solution of Eq.(9.3.3) and consider the polynomial

$$p(x) = t + \sum_{j=0}^{m} c_j x^{j+2} , \tag{9.3.4}$$

where $t$ is a parameter. We claim that for $t = 0$ , Eq.(9.3.4) is a solution of Eq.(9.2.3). To this end, we substitute Eq.(9.3.4) into Eq.(9.2.3). By collecting the coefficients of like powers of $x$ , we find that Eq(9.3.4) is a solution if and only if

$$[\,N \quad -h\,] \begin{bmatrix} c \\ t \end{bmatrix} = Nc - th = 0 , \tag{9.3.5}$$

where $h = [0, (m+2)^2, 0, \ldots , 0]^T$ . Eq.(9.3.5) is satisfied when $t = 0$ because $c$ is a solution of Eq.(9.3.3).

Therefore it follows from Lem.9.2.3 that $p$ is a solution of Eq.(9.2.2) satisfying $p(0) = 0$ . The only such solution is $p \equiv 0$ , implying $c = 0$ . $\square$

9.3.2 THEOREM. *The expression*

$$c = (N^T N)^{-1} N^T h \tag{9.3.6}$$

*is well defined and it is a continuous function of* $\alpha, \beta$ . *Furthermore, if* $\alpha, \beta$ *are the solutions of Eq.(9.2.4) in the region* $0 < \alpha < \beta$ , *then* $c$ *given by Eq.(9.3.6) is the unique solution of Eq.(9.3.1), i.e., the components of* $c$ *are the coefficients of the minimax preconditioning polynomial of Lem.8.2.2.*

PROOF: $N^T N$ is invertible because the rank of $N$ is $(m+1)$ , the number of columns. Since $N$ is a continuous function of $\alpha, \beta$ , so are $N^T$ , $N^T N$ and $(N^T N)^{-1}$ .

If $\alpha, \beta$ are the solutions of Eq.(9.2.4), then Eq.(9.3.1) has at least one solution $c$ , whose components are the coefficients of the minimax polynomial of Thm.8.2.2. However, Eq.(9.3.1) can have at most one solution because the rank of $N$ is $m+1$ . Finally, any solution of Eq.(9.3.1) must be given by Eq.(9.3.6). ☐

The algorithm to calculate the coefficients of the best minimax polynomial is presented below.

9.3.3 ALGORITHM.

Step 1. Solve for $\alpha, \beta$ from Eq.(9.2.4).
Step 2. Assemble the matrix $N$ from Eq.(9.3.2).
Step 3. Solve for $c$ from Eq.(9.3.1).

Some examples of the minimax polynomial computed by this algorithm are shown in Fig.9.3.1 and Fig.9.3.2.

We make some remarks on the computational aspects of this algorithm.

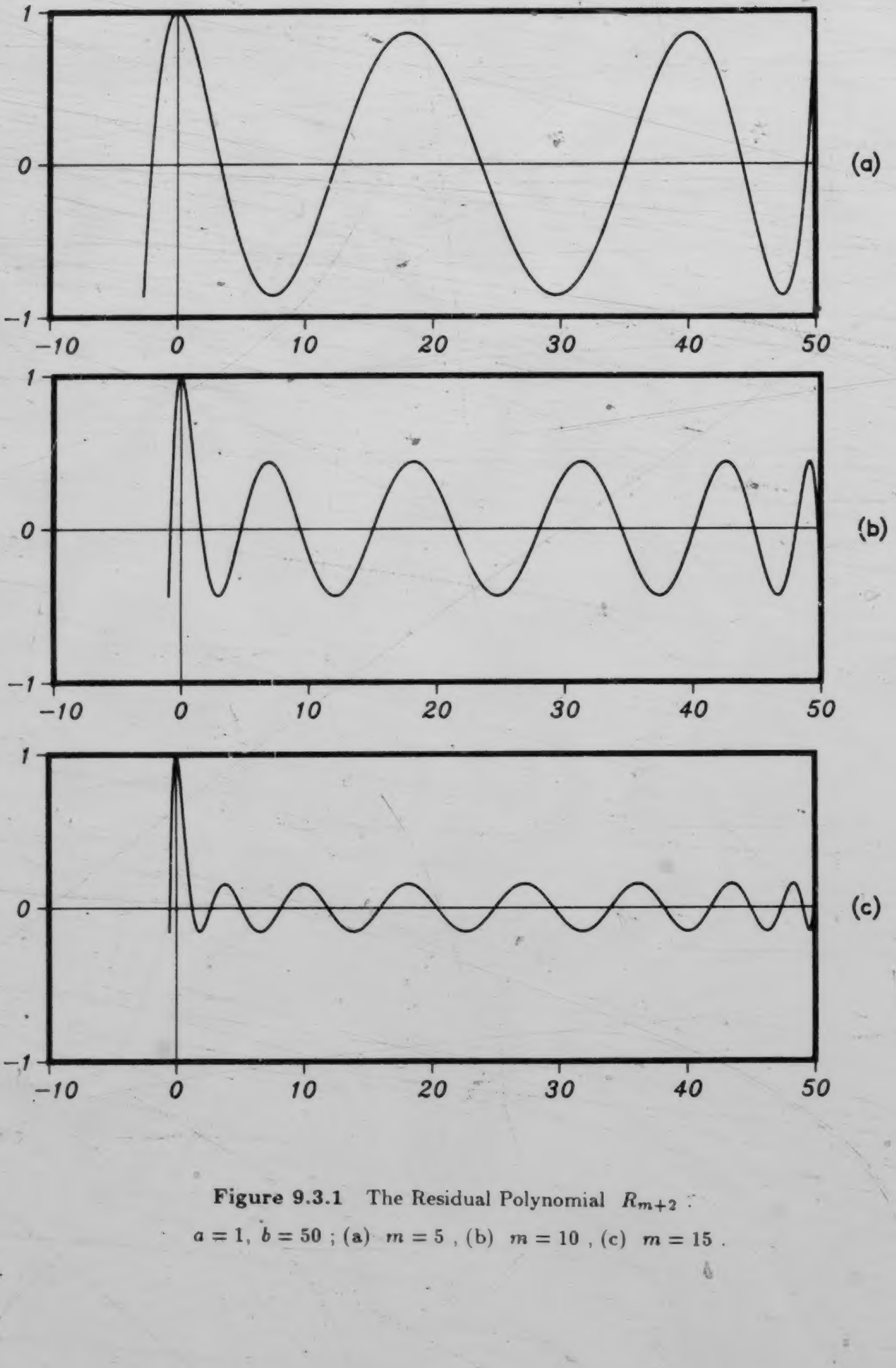

**Figure 9.3.1** The Residual Polynomial $R_{m+2}$ :
$a = 1$, $b = 50$ ; (a) $m = 5$ , (b) $m = 10$ , (c) $m = 15$ .

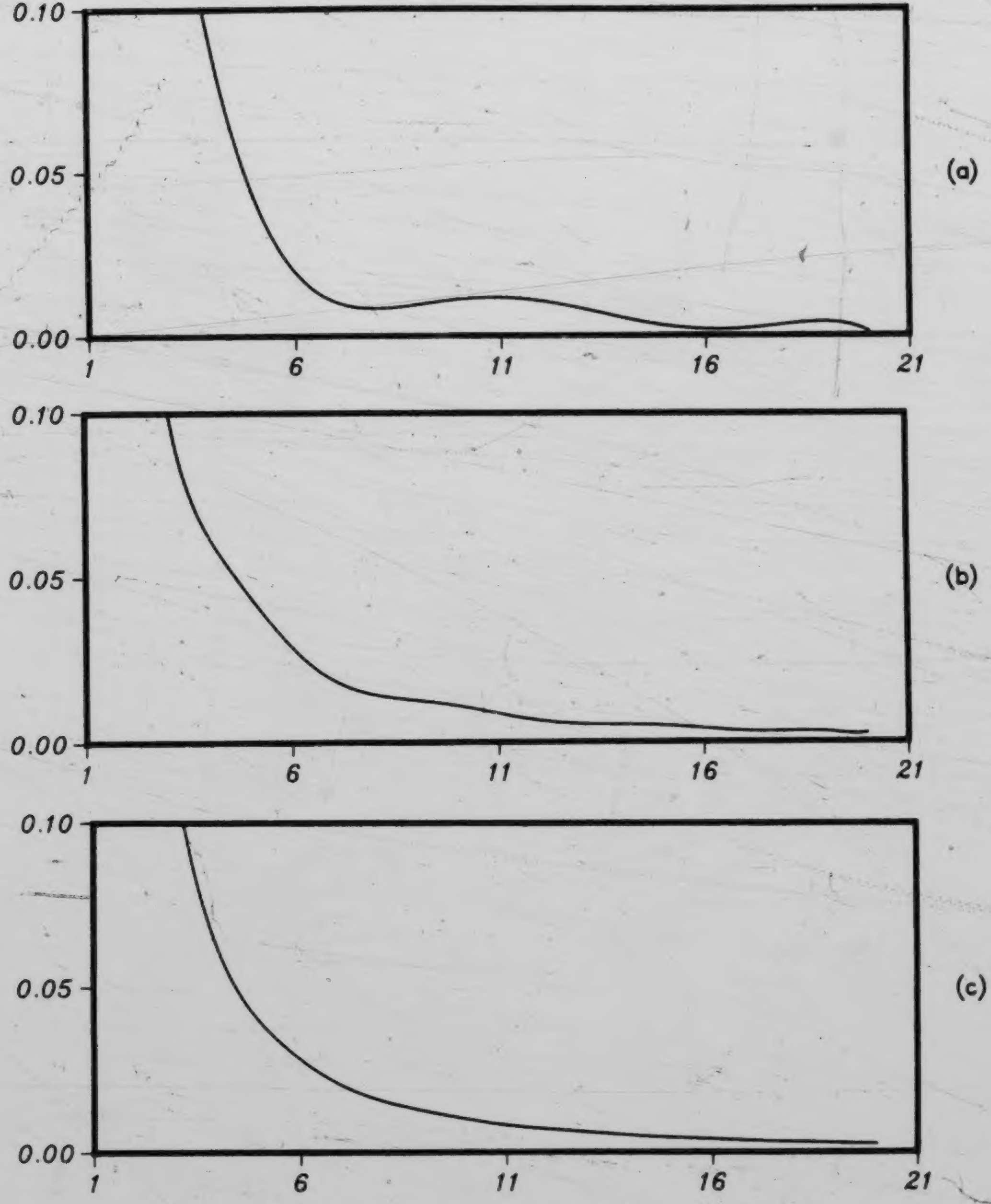

**Figure 9.3.2** The Minimax Preconditioning Polynomial $q_m$ :
$a = 1$, $b = 20$ ; (a) $m = 5$ , (b) $m = 10$ , (c) $m = 15$ .

(1) In step 1, two improper integrals need to be computed. The singularities of the integrands can be removed by some trigonometric transformations.

(2) To solve for $\alpha, \beta$, a two stage method should be used where, e.g., $\alpha = \alpha_1(\beta)$ is first computed from Eq.(9.2.4a) for given $\beta$, and it is then substituted into Eq.(9.2.4b). At each stage, the root of a scalar equation of a single variable must be found. This can be done by using Regula Falsi or bisection algorithms so that $\alpha, \beta$ can be restricted in the region $0 < \alpha < \beta$ for all computations.

(3) Even though $\alpha, \beta$ cannot be computed exactly, Eq.(9.3.1) is always solvable according to Thm.9.3.2, and, in view of the continuity guaranteed by Thm.9.3.2, the resulting coefficients $c$ are good approximations provided $\alpha, \beta$ can be obtained with enough accuracy.

(4) $c$ need not be obtained by inverting the matrix $N^T N$ which may be very ill-conditioned. In fact, Eq.(9.3.1) can be solved for $c$ by a modified Gaussian elimination method with row pivoting. At each elimination step, the search for the pivotal elements and the elimination process need to extend to all the remaining rows. The elimination steps can proceed up to row $m+1$ without failure, because the rank of $N$ is $m+1$. $c$ can be obtained by solving the first $m+1$ equations. The remaining two equations hold automatically.

# CHAPTER 10

# NUMERICAL EXPERIMENTS

A new polynomial preconditioning method to compute the solution of Eq.(8.0.1) was proposed in Chapter 8 and an efficient algorithm to compute the minimax polynomial was given in Chapter 9. In this chapter, we present some computational results by employing these methods to find iterative solutions of some linear systems.

In applications, there are perhaps very few physical problems that lead to linear systems exactly like Eq.(8.0.1). There are, however, many problems that result in linear systems of the form

$$Bu = (A^2 + E)u = b, \qquad (10.0.1)$$

where $E$ is some low rank matrix. One example is the biharmonic equation on some bounded domain. Even with an additional term $E$ in Eq.(10.0.1), the preconditioning operator $M^{-1} = q_m(A)$ proposed in Chapter 8 can still be used in PCG to solve Eq.(10.0.1). This is because the PCG algorithm does not require any additional properties of $M^{-1}$ other than that it is symmetric and positive definite. The additional term $E$, however, does effect the convergence behavior of PCG. If the effect of $E$ is small, the convergence analysis of Chapter 8 can be accepted as a reasonably good approximation of the convergence behavior of the system Eq.(10.0.1). This leads us to predict that using the polynomial preconditioning of Chapter 8 will also speed up the convergence process for the system Eq.(10.0.1). The main purpose of this chapter is to illustrate this point by numerical experiments.

In section one, we will consider a model problem of the form Eq.(8.0.1) and compare the numerical results corresponding to different methods. In section two, a linear system resulting from the discretization of a biharmonic equation is studied.

## 10.1 A Model Problem

In our model problem, $A$ is chosen as a diagonal matrix. A diagonal matrix is used not only because it is simple to compute, but also because its complete eigensystem can be artifically controlled so that an observed convergence behavior can be better understood. Of course, the solution of the system Eq.(8.0.1) can be solved directly when $A$ is a diagonal matrix, not using the PCG algorithm. However, when the PCG method is used, the solution cannot be obtained directly; the iteration proceeds as it does for any other structured matrices of the same eigensystem. This is because the PCG algorithm does not take any advantages of the explicit form of a matrix, so far as its convergence behavior is concerned. The algorithm only requires a rule that produces a vector $Ap$ for any input vector $p$. The iteration generates a Krylov space and the algorithm proceeds with the information from this Krylov space where the structure of $A$ does not play a role. In the literature, diagonal matrices are often used as test problems; see [AxLi86], [vdV87] and [vSvV86].

All the computations in this section have been carried out on an Amdahl 5870 computer, using double precision (about 14 decimal places). A diagonal matrix of dimension $500 \times 500$ is chosen as $A = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{500})$, where

$$\lambda_1 = \frac{1}{50}, \quad \lambda_j = \frac{1}{50} + (1 - \frac{1}{50})\frac{j-1}{499}, \quad j = 2, \ldots, 500.$$

The resulting system $A^2 u = b$ has a condition number

$$\kappa(A^2) = 2500.$$

The right hand side $b$ is chosen so that the exact solution is $u = (1, 1, \ldots, 1)$. The initial solution $u_0$ is composed of random numbers uniformly distributed in $[0, 1]$. The iteration process is terminated if both $\|r_j\|_\infty$ and $\|e_j\|_\infty$ are less than $10^{-7}$.

### 10.1.1 EXAMPLE.

$A^2u = b$ is solved by the approximate polynomial preconditioning $M^{-1} = q_m(A)$ where $q_m$ is computed using Algorithm 9.3.3. The number of iterations and CPU time required for convergence are given in Table 10.1.1.

**Table 10.1.1** Number of iterations and CPU time

| $m$ | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|
| Iteration | 396 | 93 | 41 | 24 | 17 | 12 | 10 | 8 | 7 | 6 |
| CPU( $10^{-6}$ sec.) | 3052 | 1231 | 702 | 501 | 422 | 347 | 331 | 297 | 283 | 271 |

### 10.1.2 EXAMPLE.

$A^2u = b$ and $Au = b$ are solved by the traditional polynomial preconditioning with $M_{A^2}^{-1} = p_m(A^2)$ and $M_A^{-1} = p_m(A)$, where $p_m$ is calculated from the Chebyshev polynomial. The number of iterations are given in Table 10.1.2.

**Table 10.1.2** Comparision of number of iterations

| $m$ | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|
| $A^2$ (TPP) | 396 | 145 | 89 | 63 | 49 | 40 | 34 | 29 | 26 | 23 |
| $A$ (TPP) | 64 | 20 | 12 | 9 | 7 | 6 | 5 | - | - | - |
| $A^2$ (APP) | 396 | 93 | 41 | 24 | 17 | 12 | 10 | 8 | 7 | 6 |

The results show that the APP preconditioning converges faster than TPP preconditioning for all tested $m$ .

## 10.2 Biharmonic Equation

Consider the Dirichlet problem for the biharmonic equation in a square $D$ with boundary $\partial D$ :

$$\Delta^2 u(x,y) = f(x,y), \qquad \text{in } D, \tag{10.2.1}$$

and $u(x,y) = g_1(x,y)$, $u_n(x,y) = g_2(x,y)$ on $\partial D$ , where $\Delta^2 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy}$ , and $u_n$ denotes the normal derivative of $u(x,y)$ .

When Eq.(10.2.1) is discretized by a 13-point finite difference scheme with a uniform mesh size $h$ , the result is a system of linear equations

$$Bu = b \tag{10.2.2}$$

where $B$ has a structure given as follows.

Define

$$R = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}_{k\times k}$$

$$T = \begin{bmatrix} 1 & & & & \\ & 0 & & & \\ & & \ddots & & \\ & & & 0 & \\ & & & & 1 \end{bmatrix}_{k\times k}.$$

Then $B = \left((I\otimes R) + (R\otimes I)\right)^2 + 2(T\otimes I) + 2(I\otimes T)$ where $\otimes$ is the standard tensor product, and $I$ is the identity. Therefore $B = A^2 + E$ where $E$ is a low rank matrix. It can be shown that both $B$ and $A$ are symmetric and positive definite matrices.

To determine the effectiveness of approximate polynomial preconditioning, consider the solution of the biharmonic equation (10.2.1) over a square region with $0 \le x, y \le 1$ . The functions $g_1(x,y)$ and $g_2(x,y)$ are chosen so that the exact solution is given by

$$u(x,y) = x^2y^2(x-1.0)^2(y-1.0)^2 . \tag{10.2.3}$$

Eq.(10.2.1) is discretized on a uniform mesh $h$ , and the total number of equations is given by $n = k^2$ , where $h = \frac{1}{k+1}$ . Numerical experiments have been performed on a CDC CYBER 205 computer with 2-pipe and 64-bit arithmetic. For the results reported in this section, the PCG iterative process is terminated when $\|r_j\|_\infty$ is less than $10^{-10}$ , where $r_j$ is the residual vector.

The performance of the preconditioned CG algorithm with two preconditionings is examined for $h = \frac{1}{100}$ . As before, by $APP$ we mean PCG with the preconditioning operator given by $M_{APP}^{-1} = \tilde{q}_m(A)$ . The polynomial $\tilde{q}_m$ is chosen as the best least squares polynomial as described in Chapter 8. $TPP$ stands for PCG with the traditional polynomial preconditioning operator $M_{TPP}^{-1} = p_m(B)$ , where $p_m$ is determined by the least squares method of Chapter 7. The number of iterations for the two methods to converge is given in the Table 10.2.1. The corresponding CPU time in seconds is given in Fig.10.2.1.

**Table 10.2.1** Number of iterations for $\mathbf{n} = 9801$

| $m$ | 0 | 1 | 2 | 3 | 4 | 5 | 10 | 20 | 30 |
|---|---|---|---|---|---|---|---|---|---|
| TPP | 1379 | 758 | 561 | 430 | 356 | 303 | 174 | 95 | 66 |
| APP | 1379 | 596 | 370 | 262 | 207 | 169 | 88 | 47 | 33 |

From Fig.10.2.1 and Table 10.2.1, it is clear that not only is the computing time significantly smaller for the $APP$ than for the $TPP$ method, but also the number of iterations for a fixed value of $m$ is smaller. The standard CG method with no preconditioning requires 5.1 seconds for convergence, whereas the best results for the $TPP$ and $APP$ methods are 3.8 and 1.2 seconds, respectively.

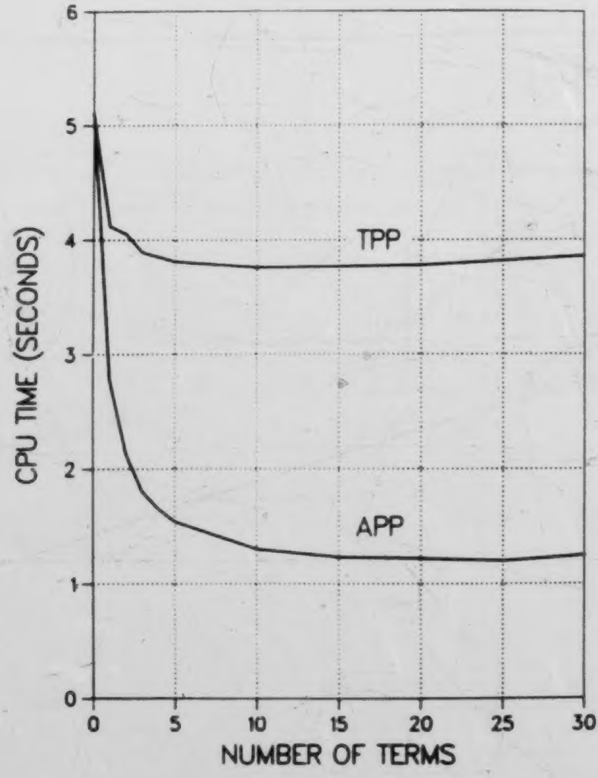

**Figure 10.2.1** CPU time

The performance of the standard CG without preconditioning and that with $APP$ preconditioning is now investigated. Table 10.2.2 compares the number of iterations required for convergence for various numbers of equations. For the $APP$ method, $m$ is fixed to be 25 for all cases tested here. The corresponding CPU time is also given in Fig.10.2.2.

**Table 10.2.2** Number of iterations

| $h^{-1}$ | 100 | 150 | 200 | 250 |
|---|---|---|---|---|
| $n$ | 9801 | 22201 | 39601 | 62001 |
| CG | 5492 | 11674 | 19568 | 31462 |
| $APP$ | 32 | 71 | 126 | 170 |

The standard CG method is not practical for solving ill-conditioned biharmonic equations due to its very slow convergence rate. However, significant improvement is clearly achieved when the $APP$ technique is used. For $n = 62001$, the computing time for $APP$ is 28.5 seconds, whereas the CG without preconditioning requires 686.1 seconds; this represents a reduction by a factor of 24.

These results confirm our earlier prediction that the approximate polynomial preconditioning technique applies not only to the system $A^2u = b$ but also to a wider class of problems encountered in applications.

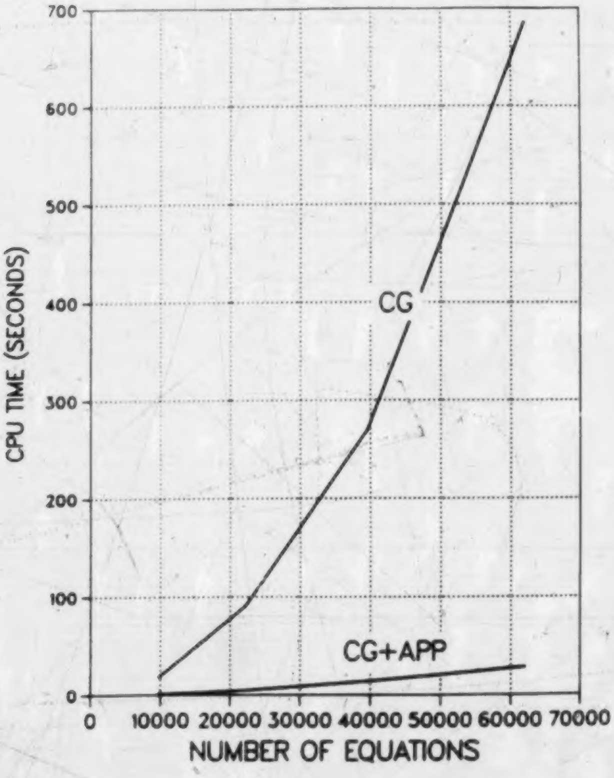

Figure 10.2.2 CPU time

# CHAPTER 11

# CONCLUSION AND DISCUSSION

In this thesis, we have investigated two important aspects of a numerical simulation process, namely, the construction of absorbing boundary conditions for hyperbolic equations and the design of a polynomial preconditioning algorithm for large linear systems of equations. In part I of the thesis, we have developed a nonreflecting principle and characterized general absorbing boundary conditions. In particular, we have shown that the optimal absorbing boundary conditions satisfy a canonical factorization form. Each factor can be tuned to annihilate a wave packet travelling at a given group velocity. The factorization form has advantages in both analysis and implementation [Higd86a]. The application of our theory also simplifies and generalizes results of [EnMa79] [EnMa80].

In part II of the thesis, we have proposed a new polynomial preconditioning method to be used in conjunction with the conjugate gradient algorithm for solving large systems of linear equations, and analyzed its convergence behavior. We have shown both theoretically and experimentally that this new polynomial preconditioning not only saves computational work but also speeds up convergence. Also in part II, we have developed an algorithm to compute a minimax polynomial which arises in applications of polynomial preconditioning.

* * *

The nonreflecting principle and the factorization theorem of part I, and the convergence analysis of part II are of theoretical interest by themselves. However, there are problems of physical interest which can make good use of the techniques of both parts of this thesis. One such example is the problem of calculation of elastic waves. In the classical theory of flexural motions, the equation governing bending of a plate is given by (see [Ache73])

$$u_{tt} + \Delta^2 u = 0\,, \tag{11.0.1}$$

where $\Delta^2 = \frac{\partial^4}{\partial x^4} + 2\frac{\partial^4}{\partial x^2 y^2} + \frac{\partial^4}{\partial y^4}$ . Eq.(11.0.1) is not hyperbolic, but its solutions can be described by dispersion waves [Whit74], so that the discussion of chapter three can also be applied here. The dispersion relation of Eq.(11.0.1) is found to be

$$\tau^2 - (\xi^2 + \eta^2)^2 = 0\,. \tag{11.0.2}$$

If only band limited solutions [ViBo82] are considered, the group velocity of wave solutions of Eq.(11.0.1) is finite. Thus when the theory of chapter four is applied, the absorbing boundary operators on the left hand side boundary $x = 0$ are given by

$$Q_d = \prod_{j=1}^{d}\Big(2\frac{\partial^3}{\partial x^3} + 2\frac{\partial^3}{\partial x \partial y^2} - v_j\frac{\partial}{\partial t}\Big)\,, \quad v_j < 0\,. \tag{11.0.3}$$

Each factor annihilates the wave packet moving with group velocity $v_j$ . It can be verified that Eq.(11.0.1) with the boundary operators Eq.(11.0.3) (on the left hand side boundary) is well-posed in the sense of Hersh.

After proper boundary conditions are formulated, Eq.(11.0.1) can be discretized by a central differencing scheme with a uniform mesh size $\Delta x = \Delta y$ . If an explicit scheme is used, then the stability requires a small time step size $\Delta t$ , which is not efficient for long time integration. Alternatively, an implicit scheme may be used. If such a scheme is chosen, at each time level, the grid solution satisfies a linear system

$$\Big(A^2 + (\frac{\Delta x}{\lambda})^2 I + E'\Big)u^{n+1} = b^n\,, \tag{11.0.4}$$

where $I$ is the identity and $A = R \otimes I + I \otimes R$ , ( $R$ is defined in § 10.2), $\lambda =$

$\frac{\Delta t}{\Delta x}$, and $E'$ is some low rank matrix dependent on the choice of the boundary conditions. When the mesh size $\Delta x$ is small, the matrix $E = (\frac{\Delta x}{\lambda})^2 I + E'$ has small effect and the approximate polynomial preconditioning technique of part II is effective for solving Eq.(11.0.4).

* * *

Absorbing boundary conditions for equations with variable coefficients or nonlinear equations are not studied in this thesis. For such problems, our approach is the frozen coefficient technique, where the variable coefficients or the nonlinear terms are treated as local constants outside the computational domain. This approach is different from prescribing the asymptotic solution at the computational boundaries, because in the former approach, solutions are allowed to oscillate outside the computational domain, but are governed by a simpler equation (an equation with constant coefficients outside the domain of computation). In fact, the frozen coefficient technique is theoretically justifiable in the case of linear equations with *slowly varying coefficients* (their derivatives are small compared to the coefficients themselves).

For equations with slowly varying coefficients, the wave propagation of solutions is described by ray theory [Whit74] [Ligh78]. Asymptotically, the energy of solutions propagates at group velocity which is calculated by freezing the coefficients. In addition, the theory of Kreiss for well-posed initial boundary value problems and the UKC (Def.3.3.2) also hold for slowly varying linear equations [Krei70] [Higd86b].

The effects of nonlinearity on the absorption properties of the absorbing boundary conditions of chapter four are under investigation. We are currently studying the absorbing boundary conditions 5.3.3 applied to the nonlinear transonic small disturbance equation Eq.(5.2.1). We have implemented three boundary conditions of 5.3.3 on a NASA Ames computer code ATRAN2. This code [BaGo77] solves the implicit finite difference approximation of Eq.(5.2.1). By calculating, under different boundary conditions, the lift and the moment coefficients over an airfoil with an oscillating trailing-edge flap, our preliminary study shows that reflections are significantly reduced by using the first-order absorbing boundary conditions, but there is no visible improvement of results by changing the first-order conditions to the second-order ones. When the first-order or the second-order absorbing boundary conditions are implemented onto the ATRAN2 code, the computational domain is substantially reduced. A detailed study of this problem is still underway.

* * *

In part II of the thesis, the approximate polynomial preconditioning of chapter 8 is applied to the system

$$A^2 u = b . \tag{11.0.5}$$

However, a similar polynomial preconditioning can be used for the solution of the system $A^k u = b$ for any integer $k > 0$. In fact, if we choose the preconditioning operator as $M^{-1} = q_m^k(A)$ and let $R_{m+k}(x) = 1 - x^k q_m(x)$, then the minimax preconditioning polynomial can be determined by minimizing the value $\max |R_{m+k}|$, and the computation of $R_{m+k}$ can be carried out as in chapter 9.

When $A$ is symmetric and nonsingular (not necessarily positive definite), $A^2$ is symmetric and positive definite, so that the CG algorithm can be applied to Eq.(11.0.5). This case is not covered in this thesis. Yet, so long as $A$ is symmetric and nonsingular, a polynomial $q_m$ can be found such that $q_m(A)$ is symmetric and positive definite. Hence Eq.(11.0.5) can be preconditioned by

$M^{-1} = q_m(A)$ in the PCG procedure. However, the algorithm of chapter 9 cannot be used directly to compute the residual polynomial $R_{m+2}$ , although the Remez algorithm is always applicable. We have not studied whether the results of chapter 9 can be generalized to the case when $A$ is indefinite.

# REFERENCES

[Ache73] J.D. Achenbach, Wave propagation in elastic solids, (North-Holland, New York, 1973).

[Adam85] L. Adams, "M-step preconditioned conjugate gradient methods", *SIAM J. Sci. Stat. Comp.*, **6**, pp.452-463 (1985).

[Ashb87] S.F. Ashby, "Polynomial preconditioning for conjugate gradient methods", *Ph.D. Dissertation*, (University of Illinois at Urbana-Champaign, 1987)

[AxLi86] O. Axelsson and G. Lindskog, "On the rate of convergence of the preconditioned conjugate gradient method", *Numer. Math.*, **48**, pp.499-523 (1986).

[BaGo77] W.F. Ballhaus and P.M. Goorjan, "Implicit finite difference computations of unsteady transonic flows about airfoils including the treatment of irreqular shock wave motions", *AIAA Paper*, pp.77-205 (1977).

[BaBr80] W.F. Ballhaus and J.O. Bridgeman, "Numerical solution techniques for unsteady transonic aerodynamics problems", Special course on unsteady aerodynamics, *AGARD Report No. 679*, (1980).

[BEHJ88a] A. Bamberger, B. Engquist, L. Halpern and P. Joly, "Parabolic wave equation approximations in heterogeneous media", *SIAM J. Appl. Math.*, **48**, pp.99-128 (1988).

[BEHJ88b] A. Bamberger, B. Engquist, L. Halpern and P. Joly, "Higher order paraxial approximations in heterogeneous media", *SIAM J. Appl. Math.*, **48**, pp.129-154 (1988).

[BaTu80] A. Bayliss and E. Turkel, "Radiation boundary conditions for wave-like equations", *Commun. Pure Appl. Math.*, **33**, pp.707-723 (1980).

[Bern26] S. Bernstein, Leçons sur les propriétés extrémales et la meilleure approximation des fonctions analytiques d'une variable réelle, (Gauthier-Villars et $C^{ie}$ , Paris, 1926), p.22

[CKT88] T.F. Chan, C.C.J. Kuo and C. Tong, "Parallel elliptic preconditioners: Fourier analysis and performance on the connection machine", presented at *Workshop on practical iterative methods for large scale computations*, (Minnesota Supercomputer Institute, Oct. 23-25, 1988).

[Chen66] E.W. Cheney, Introduction to Approximation Theory, (McGraw-Hill, New York, 1966).

[CoCo86] J.D. Cole and L.P. Cook, Transonic aerodynamics, (North-Holland, New York, 1986).

[CGO76] P. Concus, G.H. Golub and D.P. O'Leary, "A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations", Sparse matrix computations, J.R. Bunch and D.J. Rose, eds. (Academic Press, New York, 1976) pp.309-332.

[DGR79] M. Dubois, A. Greenbaum and G. Rodrigue, "Approximate the inverse of a matrix for use in iterative algorithms on vector processors", *Computing*, **22**, pp.257-268 (1979).

[Duff76] G.F.D. Duff, "Hyperbolic differential equations and waves", Boundary value problems for linear evolution partial differential equations, H.G. Garnir ed., (D. Reidel, Boston, 1976) pp.27-155.

[EnMa77] B. Engquist and A. Majda, "Absorbing boundary conditions for the numerical simulation of waves", *Math. Comput.*, **31**, pp.629-651 (1977).

[EnMa79] B. Engquist and A. Majda, "Radiation boundary conditions for acoustic and elastic wave calculations", *Commun. Pure Appl. Math.*, **32**, pp.313-357 (1979).

[EnMa80] B. Engquist and A. Majda, "Numerical radiation boundary conditions for unsteady transonic flow", *J. Comput. Phys.*, **40**, pp.91-103 (1981).

[Gård50] L. Gårding, "Linear hyperbolic partial differential equations with constant coefficients", *Acta Math.*, **85**, (1950).

[GrOr77] C. Grosch and S. Orszag, "Numerical solution of problems in unbounded regions: coordinate transforms", *J. Comput. Phys.*, **25**, pp.273-295 (1977).

[Gust78] I. Gustafsson, "A class of first order factorization methods", *BIT*, **18**, pp.142-156 (1978).

[Gust88] B. Gustafsson, "Inhomogeneous conditions at open boundaries for wave propagation problems", *Appl. Numer. Math.*, **4**, pp.3-19 (1988).

[GuKr79] B. Gustafsson and H.O. Kreiss, "Boundary conditions for time dependent problems with an artificial boundary", *J. Comput. Phys.*, **30**, pp.333-351 (1979).

[GKS72] B. Gustafsson, H.-O. Kreiss and A. Sundstrom, "Stability theory of difference approximations for mixed initial boundary value problems II", *Math. Comput.*, **26**, pp.649-686 (1972).

[HaHa88] T. Hagstrom and S.I. Hariharan, "Accurate boundary conditions for exterior problems in gas dynamics", *NASA Technical Memorandum 100807, ICOMP-88-2*, (1988).

[Heds79] G.W. Hedstrom, "Non-reflecting boundary conditions for nonlinear hyperbolic systems", *J. Comput. Phys.*, **30**, pp.222-237 (1979).

[Hers64] R. Hersh, "Boundary conditions for equations of evolution", *Arch. Rational Mech. Anal.*, **16**, pp.243-264 (1964).

[HeSt52] M.R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems", *J. Research of NBS*, **49**, pp.409-435 (1952).

[Higd86a] R.L. Higdon, "Absorbing boundary conditions for difference approximations to the multi-dimensional wave equation", *Math. Comput.*, **47**, pp.437-459 (1986).

[Higd86b] R.L. Higdon, "Initial-boundary value problems for linear hyperbolic systems", *SIAM Rev.*, **28**, pp.177-217 (1986).

[Higd87] R.L. Higdon, "Numerical absorbing boundary conditions for the wave equation", *Math. Comput.*, **49**, pp.65-90 (1987).

[JiWo89] H. Jiang and Y.S. Wong, "Absorbing boundary conditions for second-order hyperbolic equations", *J. Comput. Phys.*, (to appear, 1989).

[JMP83] O.G. Johnson, C.A. Micchelli and G. Paul, "Polynomial preconditioners for conjugate gradient calculations", *SIAM J. Numer. Anal.*, **20**, pp.362-376 (1983).

[KATW74] K.R. Kelly, R.M. Alford, S. Treifel and R.W. Ward, "Applications of finite difference methods to exploration seismology", Proc. Roy. Irish Acad. Conf. on Numerical Analysis, (Academic Press, London & New York, 1974), p. 7.

[Krei70] H.O. Kreiss, "Initial boundary value problems for hyperbolic systems", *Commun. Pure Appl. Math.*, **23**, pp.277-298 (1970).

[Krei72] H.O. Kreiss, "Boundary conditions for hyperbolic differential equations", Proc. Dundee conference on numerical solutions of differential equations, *Lecture Notes in Mathematics, No. 363*, pp.64-74 (1972).

[Ligh78] J. Lighthill, Waves in Fluids, (Cambridge University Press, London & New York, 1978).

[Lind75] E.L. Lindman, "Free-space boundary conditions for the time dependent wave equation", *J. Comput. Phys.*, pp.66-78 (1975).

[MevV77] J.A. Meijerink and H.A. van der Vorst, "An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix", *Math. Comp.*, **31**, pp.148-162 (1977).

[Newm64] D.J. Newman, "Rational approximation to $|x|$ ", *Michigan Math. J.*, **11**, pp.11-14 (1964).

[Nire73] L. Nirenberg, Lectures on linear partial differential equations, *C.B.M.S Regional Conf. Ser. in Math., no. 17*, (Amer. Math. Soc., Providence, R.I., 1973).

[Pete83] B. Petersen, Introduction to the Fourier transform & pseudo-differential operators, (Pitman Advanced Publishing Program, Boston, 1983).

[Roac72] P.J. Roache, Computational fluid dynamics, (Hermosa Press, Albuquerque, N.M., 1972).

[Saad85] Y. Saad, "Practical use of polynomial preconditionings for the conjugate gradient method", *SIAM J. Sci. Stat. Comp.*, **6**, pp.865-881 (1985).

[Saka82] R. Sakamoto, Hyperbolic boundary value problems, (Cambridge university press, 1982).

[Shil68] G.E. Shilov, Generalized functions and partial differential equations, (New York, 1968).

[Smit74] W.D. Smith, "A nonreflecting plane boundary for wave propagation problems", *J. Comput. Phys.*, **15**, pp.492-503 (1974).

[Tayl81] M.E. Taylor, Pseudodifferential operators, (Princeton Univ. Press, Princeton, NJ, 1981).

[Thom87] K.W. Thompson, "Time dependent boundary conditions for hyperbolic systems", *J. Comput. Phys.*, **68**, pp.1-24 (1987).

[Tref82] L.N. Trefethen, "Wave propagation and stability for finite difference schemes", *Ph.D. Dissertation*, (Stanford University, 1982).

[TrHa86] L.N. Trefethen and L. Halpern, "Well-posedness of one-way wave equations and absorbing boundary conditions", *Math. Comput.*, **47**, pp.421-435 (1986).

[vSvV86] A. van der Sluis and H.A. van der Vorst, "The rate of convergence of conjugate gradients", *Numer. Math.*, **48**, pp.543-560 (1986).

[vdV87] H.A. van der Vorst, "An iterative solution method for solving $f(A)x = b$, using Krylov subspace information obtained for the symmetric positive definite matrix $A$", *J. Comput. Appl. Math.*, **18**, pp.249-263 (1987).

[ViBo82] R. Vichnevetsky and J.B. Bowles, Fourier analysis of numerical approximations of hyperbolic equations, *SIAM Studies 5*, (Society for Industrial and Applied Mathematics, Philadelphia, 1982).

[Waga83] L. Wagatha, "Approximation of pseudodifferential operators in absorbing boundary conditions for hyperbolic equations", *Numer. Math.*, **42**, pp.51-64 (1983).

[Waga85] L. Wagatha, "On boundary conditions for the numerical simulation of wave propagation", *Appl. Numer. Math.*, **1**, pp.309-314 (1985).

[Whit74] G.B. Whitham, Linear and Nonlinear Waves, (Wiley-Interscience, New York, 1974).

[Wils82] J.C. Wilson, "Derivation of boundary conditions for the artificial boundaries associated with the solution of certain time dependent problems by Lax-Wendroff type difference schemes", *Proc. Edinburgh Math. Soc.*, **25**, pp.1-18 (1982).

[Wong88] Y.S. Wong, "Solving large elliptic difference equations on CYBER 205", *Parallel Computing*, **6**, pp.195-207 (1988).

[WoJi89] Y.S. Wong and H. Jiang, "Approximate polynomial preconditioning applied to biharmonic equations", *J. Supercomputing*, (to appear, 1989).